Game Theory and Decision Theory

Decision theory may be considered as the theory of a two-person game, in which nature takes the role of one of the players. The other known as statistician.

- A game (two-person, zero-sum) is a triplex $(\Theta, \mathcal{A}, L)$ where,

1. Θ is a nonempty set of possible states of nature, referred to as the parameter space.
2. $\mathcal{A}$ is a nonempty set of actions available to the statistician.
3. A loss function, $L(\theta, a)$ is a real valued function defined on $\Theta \times \mathcal{A}$.

A game can be interpreted as follows.

Nature choose a point $\theta \in \Theta$, and the statistician, without being informed of the choice nature has made, choose an action $a \in \mathcal{A}$. As a consequence of these two choices, the statistician loses an amount $L(\theta, a)$. The function L may take negative values. A negative loss may be interpreted as a gain.

Example ODD or EVEN.

Two contestants simultaneously put up either one or two fingers. One of the players, say player I, wins if the sum of the digits showing is odd, and the other player, Player II, wins if the sum of the digits showing is even. The winner in all cases receives in dollars the sum of the digits showing, this being paid to him by the loser.

To create a triplet $(\Theta, \mathcal{a}, L)$ out of this game we give player I the label "nature" and player II the label "statistician". Each of these players has two possible choices, so that,

$\Theta = \{1, 2\} = \mathcal{a}$, in which '1' and '2' stand for the decisions to put up one and two fingers, respectively. The loss function is given by the following table.

$\Theta \backslash \mathcal{a}$	1	2
1	-2	3
2	3	-4

Thus $L(1,1) = -2$, $L(1,2) = 3$, $L(2,1) = 3$, $L(2,2) = -4$

A Comparison of Game theory and Decision Theory

In a two-person game the two players are trying simultaneously to maximize their winnings (or to minimize their losses), whereas in decision theory nature chooses a state without this view in mind. This difference plays a role mainly in the interpretation of what is considered to be a good decision for the statistician. We donot assume that nature wins the amount $L(\theta, a)$ where θ and a are the points choosen by the playes.

Consider the game $(\Theta, \mathcal{A}, L)$ in which $\Theta = \{\theta_1, \theta_2\}$ and $\mathcal{A} = \{a_1, a_2\}$ and the loss function is given in the following table:

	a_1	a_2
θ_1	4	1
θ_2	-3	0.

In game theory, in which the player choosing a point from Θ is assumed to be 'intelligent' and his winning in the game are given by the function L, the only "rational" choice for him is θ_1. No matter what his opponent does, he will gain more if he chooses θ_1 than if he choose θ_2. Thus it is clear that the statistician should choose action a_2, instead of a_1.

Now,

Now suppose that the function L does not reflect the winnings of nature or that nature chooses a state without any clear objective in mind. Then we can no longer state categorically that the statistician will prefer to take action a_1.

It is assumed that nature choose the "true state" once and for all and that the statistician has at his disposal the possibility of gathering information on this choice by sampling or by performing an experiments.

~~Moreover~~ On the other hand, one can imagine a game between two intelligent adversaries in which one of the players has an advantage given to him by the rules of the game by which he can get some information on the choice his opponent has made before he himself has to make a decision. It turns out that the over-all problem which allows the statistician to gain information by sampling may simply be viewed as a more complex game.

Decision Function, Risk function

To give a mathematical structure to this process of information gathering, we suppose that the statistician before making a decision is allowed to look at the observed value of a random variable or vector, X, whose distribution depends on the true state of nature, θ. Throughout most of this discus the sample space, denoted by $\mathcal{X}$, is taken to be a Borel subset of a finite dimensional Euclidean space, and the probability distributions of X are supposed to be defined on the all Borel subsets, $\mathcal{B}$ (say) of $\mathcal{X}$. Thus for each $\theta \in \Theta$ there is a probability measure P_θ defined on $\mathcal{B}$, and a ~~corresponding~~ cumulative distribution $F_X(x|\theta)$, which represents the distribution of X when θ is the true value of the parameter. [If X is an n-dimensional vector, then ~~the notation for~~ $(X_1, \ldots X_n) = X$ and

$F_X(x|\theta) = F_{X_1, \ldots X_n}(x_1, \ldots x_n | \theta)$.]

A statistical decision problem or a statistical game is a game $(\Theta, \mathcal{A}, L)$ coupled with an experiment involving a random observation X whose distribution P_θ depends on the state $\theta \in \Theta$ chosen by nature.

On the basis of the outcome of the experiment $X = x$ (x is the ~~observation~~ observed value of X), the statistician chooses an action $d(x) \in \mathcal{A}$. Such a function d, which maps the sample space $\mathfrak{X}$ into $\mathcal{A}$, is an elementary strategy for the statistician. The loss is now the random quantity $L(\theta, d(X))$. The expected value of $L(\theta, d(X))$ as when θ is the true state of nature is called the risk function. $R(\theta, d) = E_\theta L(\theta, d(X)) = \int L(\theta, d(x)) dP_\theta(x)$
we take this expection to the Lebesgue integral.

Definition Any function d that maps the sample space $\mathfrak{X}$ into $\mathcal{A}$ is called a nonrandomized decision rule or a nonrandomized decision function, provided the risk function $R(\theta, d)$ exists and is finite for all $\theta \in \Theta$.
The class of all nonrandomized decision rules is denoted by D.
So, D consists of those functions d for which $L(\theta, d(x))$ is for each $\theta \in \Theta$ a Lebesgue integrable function of x.

In particular, D contains all simple functions.

Example The game of "odd or even" may be extended to a statistical decision problem. Suppose that before the game is played the player called "the statistician" is allowed to ask the player called "nature" how many fingers he intends to put up and that nature must answer truthfully with probability $3/4$. The statistician therefore observes a random variable X (the answer nature gives) taking values 1 or 2. If $\theta = 1$ is the true state of nature, the probability that $X=1$ is $3/4$; therefore, $P_1(X=2) = 1/4$. Similarly $P_2(X=1) = 1/4$. and $P_2(X=2) = 3/4$.

There are exactly four possible functions from $\mathcal{X} = \{1,2\}$ into $\mathcal{A} = \{1,2\}$. These are four decision rules:

$$d_1(1) = 1, \quad d_1(2) = 1$$
$$d_2(1) = 1, \quad d_2(2) = 2$$
$$d_3(1) = 2, \quad d_3(2) = 1$$
$$d_4(1) = 2, \quad d_4(2) = 2.$$

Rules d_1 and d_4 ignore the value of X. Rule d_2 reflects the belief of the statistician that nature is telling the truth, and rule d_3, that nature is not telling the truth. The risk table is as follows:

$\Theta \backslash D$	d_1	d_2	d_3	d_4
1	-2	$-3/4$	$7/4$	3
2	3	$-9/4$	$5/4$	-1

Note that the choice of a decision function should depend only on the risk function $R(\theta, d)$ (the smaller in value the better) and not otherwise on the distribution of the random variable $L(\theta, d(x))$.

Also note that, the original game $(\Theta, \mathcal{A}, L)$ has been replaced by a new game (Θ, D, R), in which the space D and the function R have an underlying structure, depending on $\mathcal{A}$, L, and the distribution of X.

One can apply decision theory in sequential analysis. For example, in sequential analysis the statistician may take observations one at a time, paying c units each time he does so. Therefore a decision rule will have to tell him both when to stop taking observations and what action to take once he has stopped. He will try to choose a decision rule that will minimize in some sense his new risk, which is defined now as the expected value of the loss plus the cost.

Some special cases

1. $\mathcal{A}$ consists of two points $\mathcal{A} = \{a_1, a_2\}$

Decision theoretic problems in which $\mathcal{A}$ consists of exactly two points are called problems in testing hypothesis. Consider the special case in which Θ is the real line and suppose that the loss function is for some fixed number θ_0 given by the formulas

$$L(\theta, a_1) = \begin{cases} l_1 & \text{if } \theta > \theta_0 \\ 0 & \text{if } \theta \leq \theta_0 \end{cases}$$

and
$$L(\theta, a_2) = \begin{cases} 0 & \text{if } \theta > \theta_0 \\ l_2 & \text{if } \theta \leq \theta_0 . \end{cases}$$

where l_1 and l_2 are positive numbers. Here we would like to take action a_1 if $\theta \leq \theta_0$ and action a_2 if $\theta > \theta_0$. The space D of decision rules consists of those functions d from the sample space into $\{a_1, a_2\}$ with the property that $P_\theta\{d(X) = a_1\}$ is well-defined for all values of $\theta \in \Theta$. The risk function is

$$R(\theta, d) = \begin{cases} l_1 P_\theta\{d(X) = a_1\} & \text{if } \theta > \theta_0 . \\ l_2 P_\theta\{d(X) = a_2\} & \text{if } \theta \leq \theta_0 \end{cases}$$

In this way probabilities of making two types of error are involved.

2. $\mathcal{A}$ consists of k points, $\{a_1, a_2, \ldots a_k\}$, $k \geq 3$.
These decision theoretic problems are called multiple decision problems. As an example, ~~where~~ an experimenter is to judge which of two treatments has a greater yield on the basis of an experiment. He may (a) decide treatment 1 is better, or (b) decide treatment 2 is better, (c) withhold judgment until more data are available, In this example $k = 3$.

3. $\mathcal{A}$ consists of the real line, $\mathcal{A} = (-\infty, \infty)$.
Such decision theoretic problems are reffered to as point estimation of a real parameter. Consider the special case in which Θ is also the real line and suppose that the loss function is given by $L(\theta, a) = c(\theta - a)^2$ where c is some positive constant. A decision function, d, in this case a real-valued function defined on the sample space, may be considered as an "estimate" of the true unknown state of nature θ. It is the statistician desire to choose the function d to minimize the risk function $R(\theta, d) = c\, E_\theta (\theta - d(x))^2$
Note that the criterion arrived at here that of choosing an estimate with a small mean square error.

Scanned with CamScanner

Randomization

Consider a game $(\Theta, \mathcal{A}, L)$ in which Θ consists of two elements $\{\theta_1, \theta_2\}$ and $\mathcal{A}$ consists of three elements $\{a_1, a_2, a_3\}$, and suppose that the loss function (or the negative of the gain) is given by the following table

	a_1	a_2	a_3
θ_1	4	1	3
θ_2	1	4	3

If nature chooses θ_1, action a_3 is preferable to action a_1. If, on the other hand, nature choose θ_2, action a_3 is preferable to action a_2. Thus a_3 is preferred to either of the other actions under the proper circumstances. However, suppose the statistician flips a fair coin to choose between actions a_1 and a_2; that is, suppose the statistician's decision is to choose a_1 if the coin comes up heads and a_2 if the coin comes up tails. This decision, denoted by δ, is a randomized decision, such decisions allow the actual choices of the action in $\mathcal{A}$ to be left to a random mechanism and the statistician chooses only the probabilities of the various outcomes. In game theory δ would be called a mixed strategy. The randomized decision δ choose action a_1 with probability $\frac{1}{2}$, action a_2 with probability $\frac{1}{2}$, and action a_3 with probability zero. The expected loss of δ is

$\frac{1}{2} L(\theta_1, a_1) + \frac{1}{2} L(\theta_1, a_2) + 0 \cdot L(\theta_1, a_3) = 5/2$

if θ_1 is true.

and if θ_2 is true then also $5/2$.

Hence δ is ~~p~~ to be preferred to a_3, for no matter what the true state of nature, the expected loss is smaller if we use δ than if we use a_3.

Moreover, any randomized decision that gives positive probability to a_3 can be improved by distributing that probability equally between the other two actions.

__Exercise__ Calculate the expected loss for δ_1 and δ_2. and show that expected loss of δ_2 is smaller than that of δ_1. where δ_1 chooses a_1, a_2, a_3 with probability p_1, p_2, p_3 where $p_3 > 0$ and ~~δ~~. $p_1 + p_2 + p_3 = 1$. δ_2 chooses a_1, a_2, a_3 with probability $p_1 + \frac{p_3}{2}$, and $p_2 + \frac{p_3}{2}$, 0 respectively.

__Remark__ In the situation in which the statistician may base his choice of action on a random variable, X, whose distribution depends on θ, again, for the same reasons, a_3 should never be chosen. Hence the statistician may disregard a_3 entirely in making his choice, thus simplifying the problem with which he is faced.

Scanned with CamScanner

(7)

More generally, a randomized decision for the statistician in a game $(\Theta, \mathcal{A}, L)$ is a probability distribution over $\mathcal{A}$. If P is a probability distribution over $\mathcal{A}$ and Z is a random variable taking values in $\mathcal{A}$, whose distribution is given by P, the expected loss in the use of the randomized decision P is

$$L(\theta, P) = EL(\theta, Z), \quad \text{provided it exists.}$$

This is to be regarded as an extension of the domain of definition of the function $L(\theta, \cdot)$ from $\mathcal{A}$ to the space of randomized decisions, for each $a \in \mathcal{A}$, shall be regarded as a probability distribution degenerate at a,

The space of randomized decisions P, for which $L(\theta, P)$ exists and is finite for all $\theta \in \Theta$ is denoted by $\mathcal{A}^*$.

With this definition, the game $(\Theta, \mathcal{A}^*, L)$ is to be considered as the game $(\Theta, \mathcal{A}, L)$ in which the statistician is allowed randomization.

By analogy, we may extend the game (Θ, D, R) to (Θ, D^*, R) where D^* is a space containing probability distributions over D. If δ denotes a probability distribution over D, $R(\theta, \delta) = ER(\theta, Z)$, where Z is a random variable taking values in D, whose distribution is given by δ.

<u>Definition</u> Any probability distribution δ on the space of nonrandomized decision functions, D, is called a randomized decision function or a randomized decision rule, provided the risk function $R(\theta, \delta) = E R(\theta, Z)$ exists and is finite for all $\theta \in \Theta$. The space of all randomized decision rules is denoted by D^*.

~~Note that, D^* contains D with an identification.~~

For simplicity we make one restriction that $\mathcal{A}^*$ contains all probability distributions giving mass one to a finite number of points of $\mathcal{A}$. Also, D^* contains all probability distributions giving mass one to a finite number of points of D.

The space D of nonrandomized decision rules may be considered as a subset of the space D^* of randomized decision rules, $D \subset D^*$ by identifying a point $d \in D$ with the probability distribution $\delta \in D^*$ degenerate at the point d.

One advantage in the extension of the definition of $L(\theta, \cdot)$ from $\mathcal{A}$ to $\mathcal{A}^*$ and the definition of $R(\theta, \cdot)$ from D to D^* is that these functions become linear on $\mathcal{A}^*$ and D^*, respectively. In otherwords, if $P_1 \in \mathcal{A}^*$, $P_2 \in \mathcal{A}^*$ and $0 \le \alpha \le 1$, then $\alpha P_1 + (1-\alpha) P_2 \in \mathcal{A}^*$ and $L(\theta, \alpha P_1 + (1-\alpha) P_2) = \alpha L(\theta, P_1) + (1-\alpha) L(\theta, P_2)$

Similarly, if $\delta_1 \in D^*$, $\delta_2 \in D^*$, and $0 \le \alpha \le 1$
then $\alpha\delta_1 + (1-\alpha)\delta_2 \in D^*$ and

$$R(\theta, \alpha\delta_1 + (1-\alpha)\delta_2) = \alpha R(\theta, \delta_1) + (1-\alpha) R(\theta, \delta_2)$$

An alternative way of setting up randomization for decision problems leads to a space of decision functions, to be denoted by $\mathcal{D}$, which is often easier to work with than D^*.

A ~~borkes~~ behavioral strategy or a behavioral decision rule, is a function, δ, which gives for each x in the sample space a probability distribution over $\mathcal{A}$, that is, an element of $\mathcal{A}^*$, $\delta(x) \in \mathcal{A}^*$. For a behavioral decision rule, δ, the risk function is now defined as

$$\hat{R}(\theta, \delta) = E_\theta\, L(\theta, \delta(x))$$

<u>Definition</u> A function $\delta(x)$, from the sample space into $\mathcal{A}^*$ is called a behavioral decision function or a behavioral decision rule, provided the ~~f~~ risk function $\hat{R}(\theta, \delta)$ exists and is finite for all $\theta \in \Theta$. The space of all behavioral decision rules is denoted by $\mathcal{D}$.

So, there are two methods of randomization correspond to the two routes in the diagram.

$$\begin{tikzcd}
 & (\Theta, D, R) \arrow[r] \arrow[ddr] & (\Theta, D^*, R) \\
(\Theta, \mathcal{A}, L) \arrow[ur] \arrow[dr] & & \\
 & (\Theta, \mathcal{A}^*, L) \arrow[r] & (\Theta, \mathcal{D}, \hat{R})
\end{tikzcd}$$

In particular $\mathcal{D}$ contains all simple functions, also $\mathcal{D}$ contains as a subset the set D of function from the sample space $\mathcal{X}$ into action space $\mathcal{A}$ ~~$\mathcal{A}^*$~~ (considered as a subset of $\mathcal{A}^*$), so that both (Θ, D^*, R) and $(\Theta, \mathcal{D}, \hat{R})$ may be considered as generalization of (Θ, D, R).

~~Among we have two routes, the route~~

We have two extensions, given by the previous diagram. In the upper route the set of functions from the sample space is considered first and then the randomization of the decision rule. In the lower route the ~~pro~~ randomization being incorporated first and set of functions from the sample space afterward.

In decision theory a behavioral decision rule tells the statistician how to randomize after observing the outcome of the experiment; whereas a randomized decision rule chooses at random a decision function that tells him before observing the outcome of the experiment exactly what action to take as a result of the experiment.

It is intuitively clear that randomized decision rules are no more general than behavioral decision rules; for the randomization provided by a rule $\delta \in D^*$ may be performed after the statistician looks at the outcome of experiment, thus proving for each $x \in \mathcal{X}$, a probability distribution over $\mathcal{Q}$.

A mapping, $\delta \rightarrow \hat{\delta}$, of this sort will embed D^* in $\mathcal{D}$ in such a way that

$$R(\theta, \delta) = \hat{R}(\theta, \hat{\delta}) \quad \text{for all } \theta \in \Theta.$$

In a similar way it is clear that the game $(\Theta, \mathcal{D}^*, \hat{R})$, which allows randomization over the set $\mathcal{D}$ of behavioral decision rules, is no more general than the game $(\Theta, \mathcal{D}, \hat{R})$. However, it is not immediately clear that (Θ, D^*, R) is not less general than $(\Theta, \mathcal{D}, \hat{R})$. It can be proved that the games (Θ, D^*, R) and $(\Theta, \mathcal{D}, \hat{R})$ are equivalent in the sense that for any $\delta \in D^*$ there is a $\hat{\delta}$ in $\mathcal{D}$ for which $R(\theta, \delta) = \hat{R}(\theta, \hat{\delta})$, for all θ and conversely.

In the rest of this note, we assume that the game (Θ, D^*, R) and $(\Theta, \mathcal{D}, \hat{R})$ are equivalent but in most of the main theorems we deal with the game (Θ, D^*, R).

Example Consider the simplest nontrivial cases, $\mathcal{A} = \{a_1, a_2\}$ and $\mathfrak{X} = \{x_1, x_2\}$. The set $\mathcal{A}^*$ of probability distributions on $\mathcal{A}$ may be taken to be the closed interval $[0,1]$ with the understanding that $\pi \in \mathcal{A}^*$ represents the prob. of taking action a_1, whereas $1-\pi$ is the prob. of taking action a_2. Now D consists of four elements $D = \{d_1, d_2, d_3, d_4\}$ where

$d_1(x_1) = a_1$	$d_2(x_1) = a_1$	$d_3(x_1) = a_2$	$d_4(x_1) = a_2$
$d_1(x_2) = a_1$	$d_2(x_2) = a_2$	$d_3(x_2) = a_1$	$d_4(x_2) = a_2$

Hence we may take.

$$D^* = \left\{(p_1, p_2, p_3, p_4) : p_i \geq 0, \sum_{i=1}^{4} p_i = 1\right\}$$

with the understanding that decision rule d_1 is chosen with probability p_1, etc. On the other hand, $\mathcal{D}$ is the space of maps from $\mathfrak{X}$ into $\mathcal{A}^*$ and can be represented as the unit square $\mathcal{D} = \{(\pi_1, \pi_2) : 0 \leq \pi_1 \leq 1, 0 \leq \pi_2 \leq 1\}$ with the understanding that if x_1 is observed $\pi_1 \in \mathcal{A}^*$ is used, whereas if x_2 is observed, then $\pi_2 \in \mathcal{A}^*$ is used, $\mathcal{D}$ is two-dimensional and D^* is three dimensional.

Exercise 1. Given $(\frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{3}{10}) \in D^*$ in the preceding example, find an equivalent element of $\mathcal{D}$.

2. Given $(1/3, 1/4) \in \mathcal{D}$ in the preceding example, find an equivalent element of D^*.

3. If $\mathcal{A} = \{a_1, a_2, \ldots a_m\}$ and $\mathfrak{X} = \{x_1, \ldots x_n\}$ show that D^* is $(m^n - 1)$ dimensional, where as $\mathcal{D}$ is $n(m-1)$ dimensional.

Optimal Decision Rules

The fundamental problem of decision theory can be stated quite simply: Given a game $(\Theta, \mathcal{A}, L)$ and a random observable X whose distribution depends on $\theta \in \Theta$, what decision rule δ should the statistician use? It is a natural reaction to search for a best decision rule, a rule that has the smallest risk no matter what the true state of nature. Unfortunately, situations in which a best decision rule exists are rare and uninteresting. In general, no one action can be presumed best over all. As an example, consider the problem of estimating a real parameter θ with quadratic loss function, $L(\theta, a) = (\theta - a)^2$. If the true state of nature is θ_0, the best action the statistician can take is $a = \theta_0$ and the best decision rule is the nonrandomized decision rule $d_0(x) = \theta_0$. There is, in general, no other significantly different decision rule as good as d_0 if θ_0 is the ~~true~~ true state of nature. Yet it is clear that d_0 cannot be considered best over all, for it is not so good as $d_1(x) = \theta_1$ if θ_1 is the true state of nature. Thus a best decision rule does not exist. Any given decision rule δ can be improved at any value θ_0 for which $R(\theta_0, \delta) > 0$ by the rule $d_0(x) = \theta_0$

Since ~~the~~ a best rule usually does not exist, two general methods, which been proposed for arriving at a decision rule, are frequently ~~sto~~ satisfactory.

Method 1. Restricting the Available Rules.

The reason a uniformly best rule usually does not exist is that there are too many others available, some of which (like d_0) are not really good. This suggests a restriction of the rules, among which the statistician is to make his choice, to a smaller class of rules, all of which have good over-all properties, in the hope that among those rules there is one that is uniformly best. Two restrictions of this kind are treated here.

1. Unbiasedness. An estimate $\hat{\theta}$ of a parameter θ is said to be unbiased if, when θ is the true value of the parameter, the mean of the distribution of $\hat{\theta}$ is θ i.e., $E_\theta \hat{\theta} = \theta \quad \forall \theta \in \Theta$

The estimate ~~in~~ d_0 of the preceding example is not unbiased, for $E_\theta d_0 = \theta_0 \quad \forall \theta \in \Theta$.

If. We apply the principle of unbiasedness and restrict the available rules to be unbiased, it is then possible that uniformly best unbiased estimate of θ will exist.

Invariance

If the decision problem is symmetric, or invariant, with respect to certain operations, then it may seem reasonable to restrict the available rules to be symmetric, or invariant, with respect to those operations also.

For example, consider the problem of estimating a parameter θ, which is a location parameter for the distribution of X, and in which the loss function is $L(\theta, a) = c(\theta - a)^2$. If the statistician is willing to estimate $a = \theta_0$ when $X = x_0$ is observed, i.e., $\hat{\theta}(x_0) = \theta_0$, then he should be willing to estimate $\theta_0 + 5$ when $X = x_0 + 5$ is observed; for his original problem is the same as the problem of estimating an unknown parameter $\varphi = \theta - 5$, which is a location parameter for the distribution $Y = X - 5$, when the loss function is $L(\varphi, b) = c(\varphi - b)^2$ ~~where b = a - 5~~. This problem is the same as the original problem, so that the statistician should be willing to estimate $b = \theta_0$ when $Y = x_0$ is observed, i.e., to estimate $a = \theta_0 + 5$ when $X = x_0 + 5$ is observed.

Any estimate $\hat{\theta}(x)$ for which $\hat{\theta}(x + c) = \hat{\theta}(x) + c$ is called an invariant estimate of θ.

Such an estimate is completely determined when the value of the estimate is given at one point. Then a uniformly best invariant estimate of θ will exist in this situation.

Method 2. Ordering the Decision Rules.

Two important principles of ordering the ~~deci~~ decision rules by which a statistician chose a decision rule are the Bayes Principle and the minimax principle.

1. The Bayes Principle.

The Bayes principle involves the notion of a distribution on the parameter space Θ called a prior distribution. Two things are ~~a~~ ~~important~~ needed of a prior distribution τ on Θ. Firstly, $ER(T, \delta)$, ~~also~~ should be ~~finite~~ exist and finite, where T is a random variable over Θ having distribution τ. $ER(T, \delta)$ if exist, is known as Bayes risk of a ~~distribution~~ decision rule δ with ~~re~~ respect to a prior distribution τ, and, we denote Bayes risk as $r(\tau, \delta)$. Secondly, we need to be able to speak of the joint distribution of T and X and of the conditional distribution of T, given X, the latter being called the posterior distribution of the parameter given the observations. It is clear ~~to~~ that ~~exist~~ any finite distribution τ on Θ ~~sati~~ satisfies these two conditions. We take the space of prior distributions, to be denoted by Θ^*, as this set of finite distributions on Θ. ~~In this rule~~ We can identify Θ as a subset of Θ^*.

In using the Bayes principle, the statistician prefers a rule δ_1 to a rule δ_2 if δ_1 has smaller Bayes risk for a fixed distribution $\tau \in \Theta^*$. This sets up a linear ordering on the space of decision rules.

Definition (Baye's decision rule)

A decision rule δ_0 is said to be Bayes with respect to the prior distribution $\tau \in \Theta^*$ if

$$r(\tau, \delta_0) = \inf_{\delta \in D^*} r(\tau, \delta).$$

The value on the right side is known as the minimum Bayes risk. Bayes rules may not exist even if the minimum Bayes risk is defined and finite. In such a case the statistician has to be satisfied with a rule whose Bayes risk is close to the minimum value.

Definition (ϵ - Bayes rules)

Let $\epsilon > 0$, A decision rule δ_0 is said to be ϵ-Bayes with respect to the prior distribution $\tau \in \Theta^*$ if

$$r(\tau, \delta_0) \leq \inf_{\delta \in D^*} r(\tau, \delta) + \epsilon$$

The prior distribution is viewed merely as a reflection of the belief of the statistician about where the true state of nature lies. The statistician naturally changes his belief after he acquires new information on the true value of the parameter by means of an experiment. His new belief should correspond to the posterior distribution of the parameter, given the outcomes.

2. The Minimax Principle

An essentially different ~~different~~ type of ordering of the decision rules may be obtained by ordering the rules according to the worst that could happen to the statistician. In other words, a rule δ_1 is preferred to a rule δ_2 if $\sup_\theta R(\theta, \delta_1) < \sup_\theta R(\theta, \delta_2)$. This relation leads to a linear ording of the space D^* of decision rules. A rule that is most preferred in this ordering is called a minimax decision rule.

<u>Definition</u> A decision rule δ_0 is said to be minimax if $\sup_{\theta \in \Theta} R(\theta, \delta_0) = \inf_{\delta \in D^*} \sup_{\theta \in \Theta} R(\theta, \delta)$

The value on the right side of the above is called the minimax value or upper value of the game.

<u>Exercise</u> Show that a decision rule δ_0 is minimax if and only if $R(\theta', \delta_0) \leq \sup_{\theta \in \Theta} R(\theta, \delta)$ for all $\theta' \in \Theta$ and $\delta \in D^*$.

<u>Note</u> Even if the minimax value is finite, there may not be a minimax decision rule, so that the statistician may have be satisfied with a rule whose maximum risk is within ϵ of the minimax value.

Definition Let $\epsilon > 0$. A decision rule δ_0 is said to be ϵ-minimax if $\sup_\theta R(\theta, \delta_0) \leq \inf_\delta \sup_\theta R(\theta, \delta) + \epsilon$.

More simply, if for all $\theta' \in \Theta$ and $\delta \in D^*$

$$R(\theta', \delta_0) \leq \sup_\theta R(\theta, \delta) + \epsilon.$$

Considering a statistical game (Θ, D, R) symmetrically, we can also defined a minimax rule for the other player, nature. Such a rule is an element of the space Θ^* of prior distributions.

Definition A distribution $\tau_0 \in \Theta^*$ is said to be least favorable if $\inf_\delta r(\tau_0, \delta) = \sup_\tau \inf_\delta r(\tau, \delta)$ ---(*)

The value on the right side of above is called the maximin or lower value of the game.

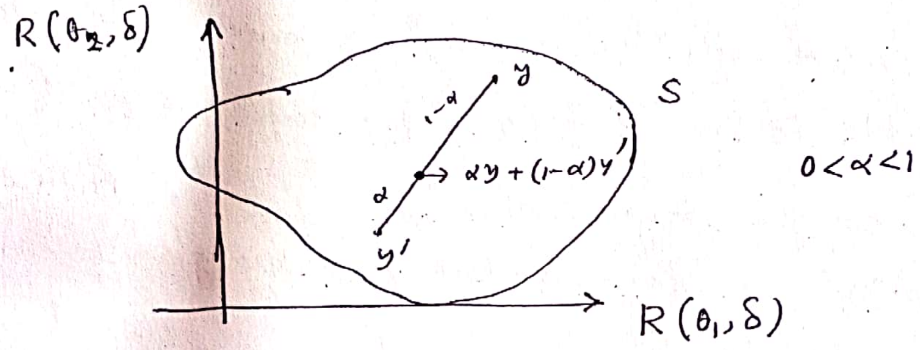
Again, there may not be a least favorable distribution. One of the advantages of allowing Θ^* to contain some continuous distributions is that a continuous distribution may then be least favorable. The name "least favorable" derives from the fact that if the statistician were told which prior distribution nature was using he would like least to be told a distribution τ_0 satisfying (*).

Exercise Show that a prior distribution τ_0 is least favorable iff $r(\tau_0, \delta') \geq \inf_\delta r(\tau, \delta)$ for all $\delta' \in D^*$ and all $\tau \in \Theta^*$.

Geometric Interpretation for finite Θ

Suppose that Θ consists of K points $\{\theta_1, \theta_2, \ldots \theta_k\}$, i.e., $\Theta = \{\theta_1, \theta_2, \ldots \theta_K\}$ and consider the set S, to be called the risk set, contained in K-dimensional Euclidean space, E_K, of points of the form $(R(\theta_1, \delta), R(\theta_2, \delta), \ldots, R(\theta_K, \delta))$, where $\delta \in D^*$.

i.e., the risk set is $S = \{(y_1, y_2, \ldots y_k) : \text{for some } \delta \in D^*, \; y_j = R(\theta_j, \delta) \text{ for } j = 1, 2 \ldots k\}$



If $K = 2$, the set can be plotted in the plane.

Definition A subset A of Euclidean K-dimensional space is said to convex if, whenever $y = (y_1, \ldots y_K)$ and $y' = (y_1', \ldots, y_K')$ are elements of A, the points

$$\alpha y + (1-\alpha) y' = (\alpha y_1 + (1-\alpha) y_1', \ldots, \alpha y_K + (1-\alpha) y_K')$$

are also elements of A for $0 < \alpha < 1$.

Lemma The risk set S is a convex subset of E_K.

Proof. Let y and y' be arbitrary points of S.

$\therefore$ There exist decision rules δ and $\delta' \in D^*$ for which $y_j = R(\theta_j, \delta)$ and $y_j' = R(\theta_j, \delta')$ for $j = 1, 2, \dots K$.

Let α be an arbitrary number ~~between~~ s.t. $0 < \alpha < 1$.

Consider the decision rule δ_α, which choose a nonrandomized decision rule in D according to the distribution δ with probability α and δ' with probability $1-\alpha$. Clearly, $\delta_\alpha \in D^*$ and

$$R(\theta_j, \delta_\alpha) = \alpha R(\theta_j, \delta) + (1-\alpha) R(\theta_j, \delta')$$

for $j = 1, 2, \dots k$.

If $Z = (z_1, z_2, \dots z_K)$ where $z_j = R(\theta_j, \delta_\alpha)$

$j = 1, 2, \dots K$ then $Z = \alpha y + (1-\alpha) y' \in S$

Completing the proof.

Definition The convex hull of a set S_0 is the smallest convex set containing S_0 or, alternatively, as the intersection of all convex sets containing S_0.

It can be shown that S is the convex hull of the set S_0, where S_0 is the nonrandomized risk set, $S_0 = \{(y_1, \dots y_K) : \text{for some } d \in D,\ y_j = R(\theta_j, d) \text{ for } j = 1, \dots k\}$

Clearly, convex hull of S_0 is a subset of S.

Because the risk function contains all the information about a decision rule as far as we are concerned, the risk set S contains all the information about the decision problem.

For a given decision problem (Θ, D^*, R) for Θ finite the risk set S is convex; Conversely, for any convex set S in k-dimension there is a decision problem (Θ, D^*, R) in which Θ consists of k points, whose risk set is the set S.

Consequently, we may discuss minimax and Bayes rules for the set S without reference to the associated decision problem.

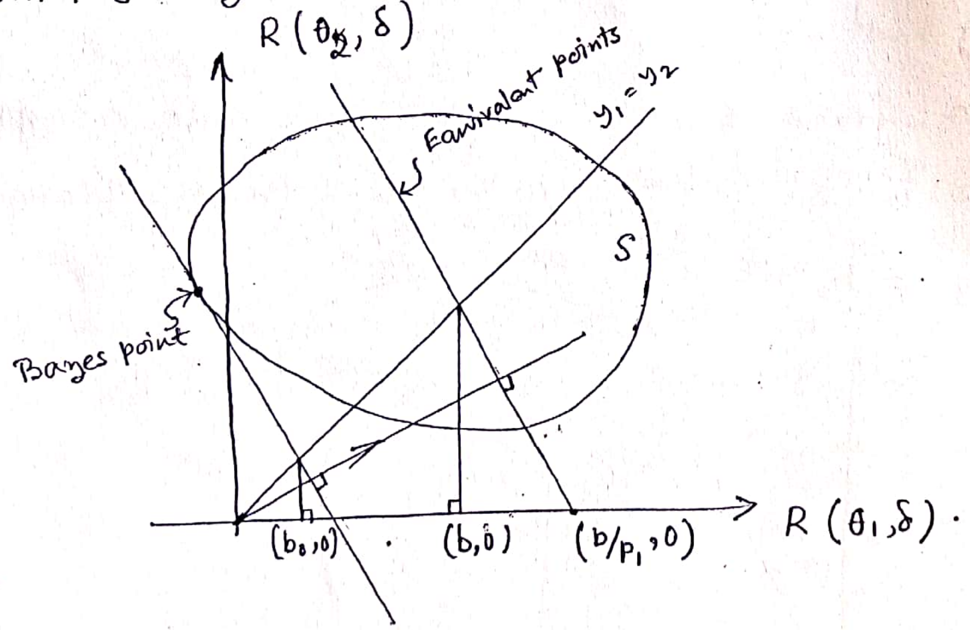
Bayes Rules

A prior distribution for nature when Θ consists of K points is merely a K-tuple of nonnegative numbers $(p_1, p_2, \dots, p_K)$ such that $\sum_{i=1}^{K} p_i = 1$, with the understanding that p_j represents the probability that nature chooses θ_j. The expected risk for the point $(y_1, y_2, \dots y_K)$ is $\sum_j p_j y_j = \sum_j p_j R(\theta_j, \delta)$ corresponds to the decision rule δ.

All the points that yield the same risk are equivalent in the ordering given by the Bayes principle for the prior distribution $(p_1, p_2, \dots, p_K)$.

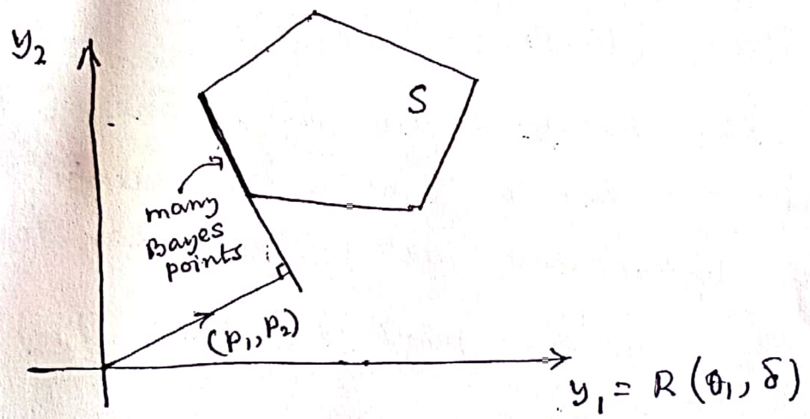
Thus all the points on the hyperplane $\sum p_j y_j = b$ for any real number b are equivalent.

Every such hyperplane is perpendicular to the vector from the origin to the point $(p_1, p_2, \dots p_K)$.

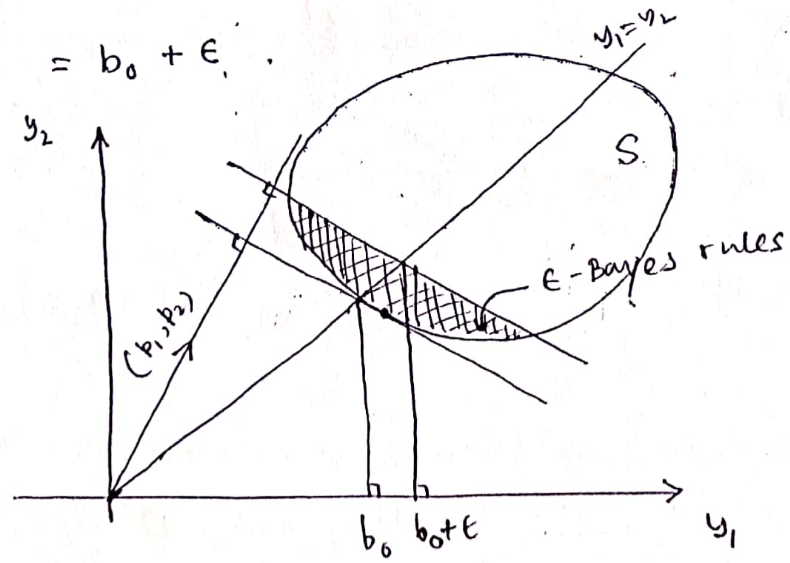


The quantity b can be visualized by noting that the point of intersection of the diagonal line $y_1 = y_2 = \dots = y_K$ with the plane $\sum p_j y_j = b$ must occur at $(b, b, \dots, b)$.

To find the Bayes rules we find the infimum of those values of b, call it b_0, for which the plane $\sum p_j y_j = b$ intersects the set S. Decision rules corresponding to points in this intersection are Bayes rules with respect to the prior distribution $(p_1, p_2, \ldots, p_k)$. Of course, Bayes rules do not exist when the set S does not contain its boundary points. On the other hand, there may be many Bayes points with respect to a given prior distn.



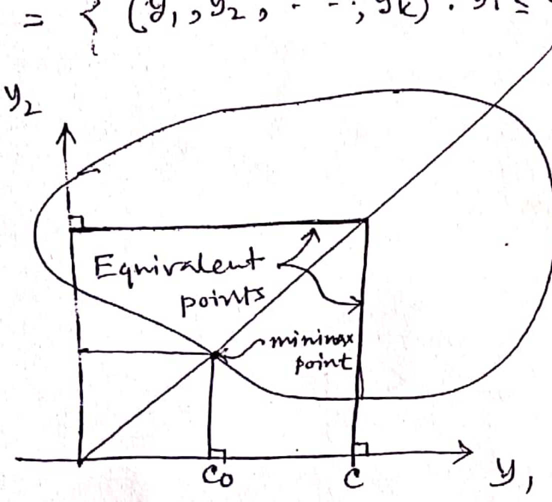
For a fixed $\epsilon > 0$, the ϵ-Bayes rule correspond to points in S which are on or below the plane $\sum p_j y_j = b_0 + \epsilon$.



Minimax Rules

The maximum risk for a fixed δ is $\max_j R(\theta_j, \delta) = \max_j y_j$. Any points $y \in S$ that give rise to the same value of $\max_j y_j$ are equivalent in the ordering given by the minimax principle.

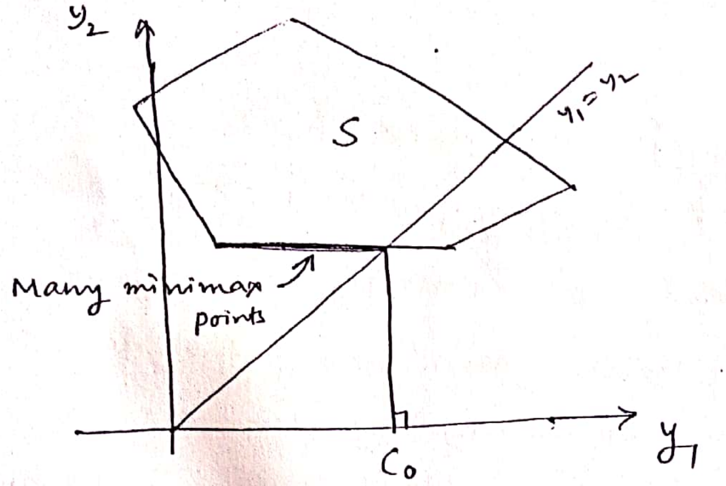
Thus all points on the boundary of the set $Q_c = \{(y_1, y_2, \cdots, y_k) : y_i \le c \text{ for } i = 1, 2, \ldots, k\}$ for any real c are equivalent.

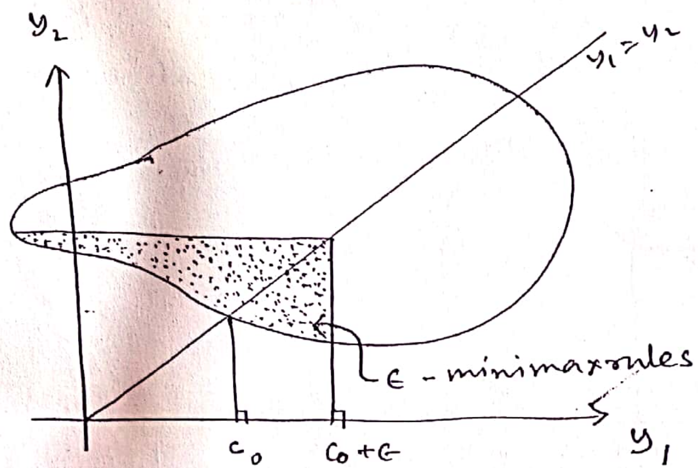


To find the minimax rules we find the infimum of those values of c, call it c_0, such that the set $Q_c \cap S \neq \phi$.

Any decision rule δ, whose associated risk point is an element of the intersection $Q_{c_0} \cap S$, is a minimax decision rule. Of course, minimax decision rules do not exist when the set S does not contain its boundary points. Again, there may be many minimax points for a given decision problem.

For a fixed $\epsilon > 0$ the rules corresponding to points in the intersection $Q_{c_0+\epsilon} \cap S$ are ϵ-minimax rules. The number c_0 is the minimax risk.

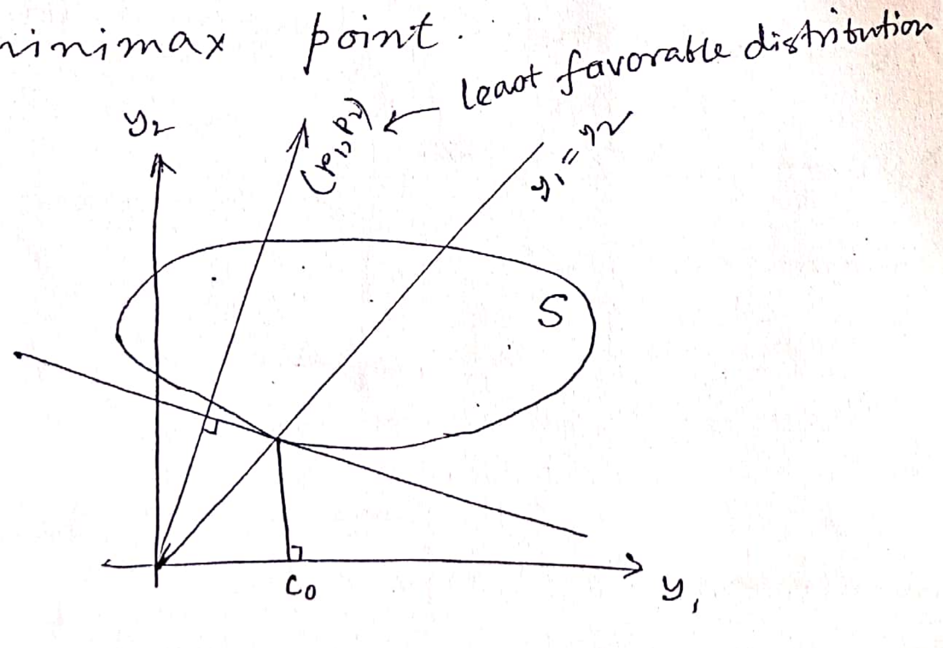




A minimax strategy for nature, which is called a least favorable distribution, may be visualized geometrically. A strategy for nature is a prior distribution, $\tau = (p_1, p_2, \cdots, p_k)$, which, as we have seen, represents the family of planes perpendicular to $(p_1, p_2, \cdots, p_k)$. In using a Bayes rule to obtain $\inf_\delta r(\tau, \delta)$, the Statistician finds the plane out of this family that is tangent to and below S. Because the minimum Bayes risk $\inf_\delta r(\tau, \delta)$ is b_0 where $(b_0, b_0, \cdots, b_0)$ is the intersection of the line $y_1 = y_2 = \cdots = y_k$ and the plane, tangent to and below S, and perpendicular to $(p_1, \cdots, p_k)$.

A least favorable distribution is the choice of $(p_1, p_2, \ldots p_k)$ that makes this intersection as far up the liner as possible. It is clear that b_0 is not greater than c_0, the minimax risk, so that if we find a prior distribution whose minimum Bayes risk is c_0 this distribution is least favorable. Clearly, if ~~Bayes~~ minimax ~~exist~~ point exist then least favorable distribution exist and the minimax risk and the Bayes risk with respect to least favorable distribution is same.

~~To find~~ the least favorable distribution, is the vector corresponding to the perpendicular ~~line~~ of the tangent plane at ~~mi~~ minimax point.



As an example, Consider the game of "odd or even". In this game $\Theta = \{\theta_1, \theta_2\}$, $\mathcal{A} = \{a_1, a_2\}$, and the loss is as follows

$L(\theta, a)$

$\theta \backslash a$	a_1	a_2
θ_1	-2	3
θ_2	3	-4

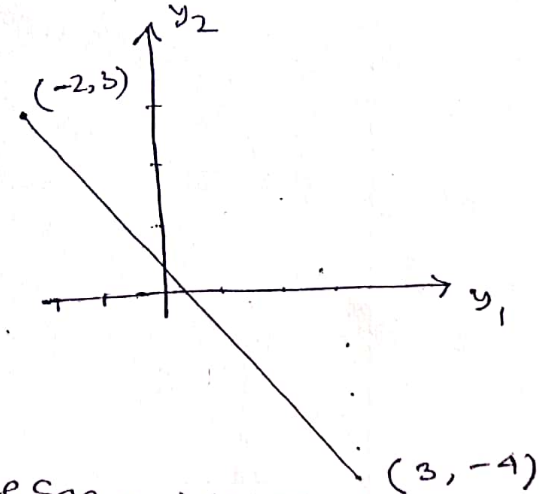
A randomized strategy $\delta \in \mathcal{A}^*$ may be represented as a number q, $0 \le q \le 1$, with the understanding that a_1 is taken with the probability q and a_2, with probability $1-q$.

Thus, the risk set for the game $(\Theta, \mathcal{A}^*, L)$ is

$$S = \{(L(\theta_1, \delta), L(\theta_2, \delta)) : \delta \in \mathcal{A}^*\} = \cancel{\{(3-5q,}$$

$$= \{(3 - 5q, -4 + 7q) : 0 \le q \le 1\}$$

which is merely the line segment joining $(3, -4)$ and $(-2, 3)$.



The minimax strategy occurs when $3 - 5q = -4 + 7q$, i.e. when $q = 7/12$ and the minimax risk is $3 - 5(7/12) = 1/12$.

This minimax rule is also a Bayes rule with respect to a prior distribution taking θ_1 with probability p and θ_2 with probability $1-p$ if the vector $(p, 1-p)$ is perpendicular to the line segment S. Because the slope of the vector $(p, 1-p)$ is $\frac{1-p}{p}$, ~~Therefor $\frac{1-p}{p} = -\frac{1}{(-7/5)}$~~

Therefore $\left(\frac{1-p}{p}\right)\left(7/5\right) = 1$, $\therefore$ $p = 7/12$

In fact, every $\delta \in \mathcal{A}^*$ is Bayes with respect to this prior distribution, which is a least favorable distribution, for its minimum Bayes risk is equal to the minimax risk, namely $1/12$.

This is a rather interesting situation: The statistician has a rule that restricts his expected loss (at most) to $1/12$, and nature has a rule that maintains the statistician's expected loss (at least) to $\frac{1}{12}$. It seems reasonable to call $\frac{1}{12}$ the value of the game.

The minimax theorem shows that this situation occurs in all finite games i.e. when Θ and $\mathcal{A}$ both finite.

Exercise 1. In the game of extended odd or even ~~gar~~ the risk function is given as

$\theta \backslash d$	d_1	d_2	d_3	d_4
θ_1	-2	$-3/4$	$7/4$	3
θ_2	3	$-9/4$	$5/4$	-4

The risk set S must contain all lines between any two of the points $(-2,3)$, $(-3/4, -9/4)$, $(7/4, 5/4)$, $(3,-4)$. Because S is convex, it is exactly the convex hull of theses four points. Find

~~Find~~ (a) the minimax rule for the statistician and the minimax risk.

(b) Find the prior distribution with respect to which the minimax rule is Bayes. Is this distribution least favorable?

(c) Find the Bayes rule with respect to the prior distribution giving weight $\frac{1}{2}$ to each of θ_1, θ_2.

2. Let $\Theta = \{0,1\}$, $\mathcal{A} = \{0,1\}$ and the loss function be $L(0,0) = L(1,1) = 0$, $L(1,0) = L(0,1) = 1$

(The statistician tries to guess what nature has chosen. If he guesses correctly he loses nothing, and vise-versa.) Suppose the statistician observes the random variable X with the discrete distribution

$P(X = x \mid \theta) = 2^{-k}$ if $x = k - \theta$ for $k = 1, 2, \ldots$

Describe the set of all nonrandomized decision rules. Plot the risk set S in the plane. Find a minimax decision rule and a least favorable distribution.

The form of Bayes rules for Estimation problems

One advantage that Bayes approach has over the minimax approach to decision theory problems is that the search for good decision rules may be restricted to the class of non-randomized decision rules, i.e, if a Bayes rule w.r.t. to a prior distribution τ exists, there exists a nonrandomized Bayes rule w.r.t. τ; similary for ϵ - Bayes rules.

Suppose that $\delta_0 \in D^*$ is Bayes w.r.t. a distribution τ over Θ. Let Z denote the random variable with values in D whose distribution is given by δ_0. Then $r(\tau, \delta_0) = E\, r(\tau, Z)$, (assumins that interchange of order of integrations ~~are~~ is valid) but, because δ_0 is Bayes w.r.t τ, $r(\tau, \delta_0) \leq r(\tau, d)$ for all $d \in D$. Therefore, $r(\tau, Z) = r(\tau, \delta_0)$ with probability 1, so that any $d \in D$ that Z chooses with probability 1 satisfy $r(\tau, d) = r(\tau, \delta_0)$ implying that d is Bayes w.r.t. τ.

Given the prior distribution τ, we want to choose a nonrandomized decision rule $d \in D$ that minimizes the Bayes risk

$$r(\tau, d) = \int R(\theta, d)\, d\tau(\theta)$$

where $R(\theta, d) = \int L(\theta, d(x))\, dF_x(x \mid \theta)$

A choice of θ by the distribution $\tau(\theta)$, followed by a choice of X from the distribution $F_X(x|\theta)$, determines a joint distribution of θ and X, which in turn can be determined by first choosing X according to its marginal distribution

$$F_X(x) = \int F_X(x|\theta)\, d\tau(\theta)$$

and then choosing θ according to the conditional distribution of θ, given $X = x$,

$$\tau(\theta|x) = \frac{f_X(x|\theta)\,\tau(\theta)}{f_X(x)}.$$

Hence by a change in the order of integration we may write

$$r(\tau, d) = \int \left[\int L(\theta, d(x))\, d\tau(\theta|x)\right] dF_X(x).$$

To find a function $d(x)$ that minimizes the above ~~in~~ double integral, we may minimize

$$\int L(\theta, d(x))\, d\tau(\theta|x).$$

In other words, the Bayes decision rule minimizes the posterior conditional expected loss, given ~~obser~~ observations.

Scanned with CamScanner

Example

Consider the estimation problem in which $\Theta = \mathcal{A} = (0, \infty)$ and $L(\theta, a) = c(\theta - a)^2$, where $c > 0$. Suppose the statistician observed the value of a random variable X having a uniform distribution on the interval $(0, \theta)$ with density

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

We are to find a Bayes rule w.r.t the prior distribution τ with density

$$g(\theta) = \begin{cases} \theta e^{-\theta} & \text{if } \theta > 0 \\ 0 & \text{if } \theta < 0 \end{cases}$$

The joint density of X and θ is

$$h(x, \theta) = g(\theta) f(x \mid \theta)$$

and the marginal distribution of X has the density

$$f(x) = \int h(x, \theta)\, d\theta = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Hence the posterior expected loss ~~given~~ $X = x$ is

$E\{L(\theta, a) \mid X = x\} = c$

Hence the posterior distribution of θ, given $X = x$, has the density

$$g(\theta | x) = \frac{h(x,\theta)}{f(x)} = \begin{cases} e^{x-\theta} & \text{if } \theta > x \\ 0 & \text{if } \theta < x \end{cases}$$

where $x > 0$. The posterior expected loss given $X = x$ is

$$E\{L(\theta,a) | X = x\} = ce^{x}\int_{x}^{\infty}(\theta - a)^2 e^{-\theta}\, d\theta$$

To find the action a that minimizes this expectation

$$\frac{d}{da} E\{L(\theta,a) | X = x\} = -2ce^{x}\int_{x}^{\infty}(\theta - a)e^{-\theta}\, d\theta = 0$$

$$\therefore \quad d(x) = a = \frac{\int_{x}^{\infty}\theta e^{-\theta}\, d\theta}{\int_{x}^{\infty} e^{-\theta}\, d\theta} = x + 1.$$

The posterior expected loss, given $X = x$, for a quadratic loss function at action a is merely the second moment about a of the posterior distribution of θ given x.

$$E\{L(\theta,a) | X = x\} = c\int (\theta - a)^2\, d\tau(\theta|x).$$

This quantity is minimized by taking a as the mean of the distribution. Hence the Bayes rule is simply

$$d(x) = E(\theta | X = x).$$

Scanned with CamScanner

Exercise

1. Consider the decision problem with $\Theta = \mathcal{A} = \mathbb{R}$ and loss function

$$L(\theta, a) = \begin{cases} k_1 |\theta - a| & \text{if } a \le \theta \\ k_2 |\theta - a| & \text{if } a > \theta, \end{cases}$$

where $k_1, k_2 > 0$.

The function $f(b) = E\,L(Z, b)$ for an arbitrary random variable Z with finite first moment is when b is a pth quantile of the distribution of Z. Find p as a function of k_1 and k_2 and state a rule for finding Bayes rules, using this loss function.

2. Let $\Theta = \mathcal{A} = (0, \infty)$ and $L(\theta, a) = c|\theta - a|$.

Let the distribution of X is given by

$$f(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

and find a Bayes rule w.r.t to the prior distribution τ of with density

$$g(\theta) = \begin{cases} \theta \cdot e^{-\theta} & \text{if } \theta > 0 \\ 0 & \text{if } \theta < 0 \end{cases}$$

Compare this rule with the Bayes rule found in example, using squared error loss.

3. Let $\Theta = \mathcal{A} \doteq \mathbb{R}$ and the loss function be

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

where $c > 0$. The function $f(b) = EL(Z, b)$ for an arbitrary random variable Z is minimized when b is the midpoint of the modal interval of length $2c$. Define "modal interval of length $2c$" so that this makes sense and state a rule for finding Bayes rules, using this loss function.

4. Let $\Theta = \mathcal{A}$ be a subset of the real line and let $L(\theta, a) = w(\theta)(\theta - a)^2$. Show that if a Bayes rule d with respect to a prior distribution τ is an unbiased estimate of θ and $E\,w(\theta) < \infty$, then $r(\tau, d) = 0$

5. Let $\Theta = (0, \infty)$, let $\mathcal{A} = \mathbb{R}$, and let

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

Let the distribution of X be Poisson with parameter $\theta > 0$, $f_X(x \mid \theta) = e^{-\theta} \dfrac{\theta^x}{x!}$, $x = 0, 1, 2 \ldots$

Take as the prior distribution of θ the gamma distribution $\mathcal{G}(\alpha, \beta)$ with density

$$g(\theta) = (\Gamma(\alpha)\beta^\alpha)^{-1} e^{-\theta/\beta} \theta^{\alpha - 1} \quad \text{for } \theta > 0.$$

where $\alpha > 0$ and $\beta > 0$.

(a) Show that the posterior distribution of θ given $X=x$ is the gamma distribution $\mathcal{G}(\alpha+x, \frac{\beta}{\beta+1})$.

(b) The first two moments of the gamma distribution $\mathcal{G}(\alpha,\beta)$ are $\alpha\beta$ and $\alpha(\alpha+1)\beta^2$. Show that the Bayes rule w.r.t $\mathcal{G}(\alpha,\beta)$ is $d_{\alpha,\beta}(x) = \frac{\beta(\alpha+x)}{\beta+1}$

(c) Show that the maximum likelihood estimate $d(x)=x$ is not a Bayes rule. Note that if $\Theta = [0,\infty)$, $d(x) = x$ will be a Bayes rule.

5. Let $\Theta = (0,1)$, $\mathcal{A} = [0,1]$, $L(\theta,a) = (\theta-a)^2$, and let the distribution of X be binomial with n trials and probability θ of success.,

$$f_X(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x=0,1,\cdots,n$$

Take as a prior distribution of θ the beta distribution $\mathcal{B}(\alpha,\beta)$ with density

$$g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $0<\theta<1$, where $\alpha,\beta>0$.

(a) Show that the posterior distribution of θ given $X=x$ is the beta distribution $\mathcal{B}(\alpha+x, \beta+n-x)$.

(b) ~~The first mome~~ Show that the Bayes rule w.r.t $\mathcal{B}(\alpha,\beta)$ is $d_{\alpha,\beta}(x) = \frac{\alpha+x}{\beta+\alpha+x}$

(c) Show that the maximum likelihood estimate of $\hat{d}(x) = x/n$ is not a Bayes rule.

(d) Consider the loss is changed to $L(\theta; a) = (\theta - a)^2 / [\theta(1-\theta)]$. Show that the maximum likelihood estimate $d(x) = x/n$ is Bayes w.r.t the uniform distribution on $(0,1)$.

7. Let X and Y be independent with binomial distributions

$$f_{X,Y}(x, y \mid p_1, p_2) = \binom{n}{x}\binom{n}{y} p_1^x p_2^y (1-p_1)^{n-x} (1-p_2)^{n-y}$$

$x = 0, 1, \ldots n$, $y = 0, 1, 2, \ldots, n$

An estimate is required of the difference $p_1 - p_2$, using square error loss

$$L((p_1, p_2), a) = (p_1 - p_2 - a)^2, \text{ where}$$

$|a| \leq 1$, $0 \leq p_1, p_2 \leq 1$.

Find the Bayes estimate w.r.t the prior distribution which assigns independent uniform distributions on $(0,1)$ to p_1 and p_2.

Scanned with CamScanner

Admissibility and Completeness

Defn (Natural ordering).

A decision rule δ_1 is said to be as good as a rule δ_2, if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$, $\forall\ \theta \in \Theta$.

A rule δ_1, is said to be better than a rule δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ $\forall\ \theta \in \Theta$, and $R(\theta, \delta_1) < R(\theta, \delta_2)$ for at least one $\theta \in \Theta$.

A rule δ_1, is said to be equivalent to a rule δ_2 if $R(\theta, \delta_1) = R(\theta, \delta_2)$ $\forall\ \theta \in \Theta$.

<u>Theorem</u> The natural ordering gives a partial ordering of the space D^* (or $\mathcal{D}$) of decision rules.

<u>Pf</u> Exercise.

<u>Important Fact</u>

In any linear orderding of ~~a partial ordering~~ the decision rules there is a minimal requirement, that the linear ordering shall not disagree with the natural ordering "as good as" that is, if δ_1 is as good as δ_2, then δ_2 is not to be preferred to δ_1.

Both the minimax and Bayes ordering statisfy this requirement.

Definition A rule δ is said to admissible if there exists no rule better than δ. A rule is ~~said~~ said to be inadmissible if it is not admissible.

Admissibility, in a very weak sense, is an optimum property. However, if the risk set S does not contain its boundary points, every rule may be inadmissible.

Definition A class C of decision rules, $C \subseteq D^*$, is said to be complete, if given any rule $\delta \in D^*$ not in C, there exists a rule $\delta_0 \in C$ that is better than δ.

A class C of decision rules is said to be essentially complete, if, given any rule δ, there exists a rule $\delta_0 \in C$ that is as good as δ.

Lemma 1. If C is a complete class, and A denotes the class of all admissible rules, then $A \subseteq C$.

Pf. Exercise

Lemma 2 If C is an essentially complete class and $\delta \notin C$, is an admissible rule, then $\exists\ \delta' \in C$ such that δ' is equivalent to δ.

Pf. Exercise.

<u>Definition</u> A class C of decision rules is said to be minimal complete if C is complete and if no proper subclass of C is complete.

A class C of decision rules is said to be minimal essentially complete if C is essentially complete and if no proper subclass of C is essentially complete.

The use of the notion of complete class is that, there ~~statistician~~ is no need for the statistician to look outside this class to find a decision rule, Thus the statistician can ~~f~~ simplify his task by finding a small complete class from which to make his choice. A minimal complete or minimal essentially complete class may not exist,

if minimal complete class exist then this is a smallest class and affords the maximum reduction of his problem.

Theorem If a minimal complete class exists, it consists of exactly the admissible rules.

Pf Let C denotes a minimal complete class and let A denote the class of all admissible rules. We are to show $C = A$. Lemma 1 implies that $A \subseteq C$. ~~Let~~ We want to show $C \subseteq A$. If possible let, $\delta_0 \in C$ s.t. $\delta_0 \notin A$. then $\exists$ $\delta_1 \in \mathcal{S}$ s.t. δ_1 is better than δ_0 ($\because \delta_0$ is inadmissible). If $\delta_1 \notin C$, then, there exists a $\delta_2 \in C$ that is better than δ, and hence better than δ_0. and if $\delta_1 \in C$, let $\delta_2 = \delta_1$.

Therefore, in either case $\exists$ a rule $\delta_2 \in C$ which is better than δ_0.

Now let $C_1 = C - \{\delta_0\}$. Then C_1 is Complete, contradicting the fact that C is minimal.

C_1 is complete because, if $\delta \notin C_1$ and $\delta \neq \delta_0$, $\exists$ $\delta' \in C$ which is better than δ, if $\delta' = \delta_0$ than $\delta_2 \in C_1$ is better than δ if $\delta' \neq \delta_0$, then $\delta' \in C_1$ and better than δ.

In any case, there exists ~~an~~ rule in C_1 better than δ, which proves the completeness of C_1

Exercise 1. Show that every complete class is essentially complete.

2. Show that if C is complete and contains no proper essentially complete subset, then C is minimal complete and minimal essentially complete.

3. Prove the following converse of the above Theorem: if the class of admissible rules is complete, it is minimal complete.

4. Find a counterexample to the following: if C_1 and C_2 are complete, then $C_1 \cap C_2$ is essentially complete.

Theorem 1. If for a given prior distribution τ a Bayes rule w.r.t. τ is unique up to equivalence, this Bayes rule is admissible.

Pf Let δ be a Bayes rule w.r.t τ and unique up to equivalence.

If posible let. δ is ~~so~~ inadmissible.

then $\exists$ δ', which is better than δ.

$\therefore$ $R(\theta, \delta') \leq R(\theta, \delta)$ $\forall$ θ

and for atleast one $\theta_0 \in \Theta$.

$R(\theta_0, \delta') < R(\theta_0, \delta)$.

$\therefore$ $r(\tau, \delta') \leq r(\tau, \delta)$

But, δ is a. Bayes rule.

$\therefore$ $r(\tau, \delta) = r(\tau, \delta')$

$\therefore$ δ' is also a Bayes rule but

$\therefore$ δ and δ' are not equivalent

a contradiction. Hence the proof.

Note

Recall that if there exists a randomized Bayes rule w.r.t τ, there exists a nonrandomized Bayes Rule w.r.t. τ, we see that the above theorem applies essentially to nonrandomized rules. Also note that in checking uniqueness we do not have to go outside the class of nonrandomized rules. Thus, if for a given prior distribution τ the Bayes rule d w.r.t τ is unique up to equivalence among the nonrandomized rules, then d is admissible.

Theorem 2

Assume that $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$ and that a Bayes rule δ_0, w.r.t. the prior distribution $(p_1, p_2, \dots p_k)$ exists. If $p_j > 0$ for $j = 1, \dots, k$, then δ_0 is admissible.

Proof Suppose that δ_0 is inadmissible; then $\exists\ \delta' \in D^*$ which is better than δ_0; that is,

$$R(\theta_j, \delta') \leq R(\theta_j, \delta_0) \quad \forall j$$

and

$$R(\theta_j, \delta') < R(\theta_j, \delta_0) \quad \text{for some } j$$

Because all $p_i > 0$, we have

$$\sum_{j=1}^{k} p_j R(\theta_j, \delta') < \sum_{j=1}^{k} p_j R(\theta_j, \delta_0),$$

a contradiction.

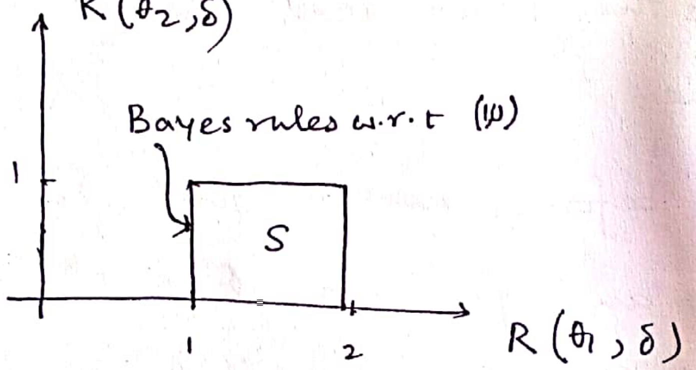
Example This counter example shows that δ_0 is not necessarily admissible if the condition $p_j > 0$ for $j = 1, 2, \dots k$ is not true in the above theorem. It also provides a counter example to theorem 1 if the Bayes rule is not unique.

Suppose that $\Theta = \{\theta_1, \theta_2\}$ and that the risk set S is the square,

$$S = \{(y_1, y_2) : 1 \leq y_1 \leq 2,\ 0 \leq y_2 \leq 1\}.$$

Consider the decision rules that Bayes w.r.t the prior distribution (p_1, p_2) with $p_1 = 1, p_2 = 0$.

Because $\sum p_j R(\theta_j, \delta) = R(\theta_1, \delta)$, it is clear that any decision rule that achieves the minimum value of $y_1 = R(\theta_1, \delta)$, namely the value $R(\theta_1, \delta) = 1$, will be a Bayes rule w.r.t. the prior distribution $(1, 0)$



Thus the rule δ_0 for which $R(\theta_1, \delta_0) = 1$, $R(\theta_2, \delta_0) = 1$ is Bayes w.r.t $(1, 0)$ (which is incidentally minimax), yet it is not admissible because the rule δ' for which $R(\theta_1, \delta') = 1$, $R(\theta_2, \delta') = 0$, is better that δ_0.

The set S can be generate as follows,

if the variable X were degenerate at zero for all θ, if $A = \{a_1, a_2, a_3, a_4\}$ and if the loss functions is given by,

	a_1	a_2	a_3	a_4
θ_1	1	1	2	2
θ_2	0	1	0	1

Here, δ_0 choose a_2 with prob. 1 and δ' choose a_1 with prob. 1.

Theorem

Let $\Theta = \mathbb{R}$ and assume that $R(\theta, \delta)$ is a continuous function of θ for all $\delta \in D^*$. If δ_0 is a Bayes rule w.r.t. a probability distribution τ on the real line, for which $r(\tau, \delta_0)$ is finite, and if the support of τ is the whole real line, then δ_0 is admissible.

Proof. As before, assume that δ_0 is not admissible and find a $\delta' \in D^*$ for which $R(\theta, \delta') \leq R(\theta, \delta_0)$ for all θ and $R(\theta_0, \delta') < R(\theta_0, \delta_0)$ for some $\theta_0 \in \mathbb{R}$. Because $R(\theta, \delta)$ is continuous in θ for all δ, there exists an $\epsilon > 0$ for which $R(\theta, \delta') \leq R(\theta, \delta_0) - \eta/2$ where $\eta = R(\theta_0, \delta_0) - R(\theta_0, \delta') > 0$, whenever $|\theta_0 - \theta| < \epsilon$. Then letting T be a ~~the~~ random variable whose distribution is τ, we have,

$$r(\tau, \delta_0) - r(\tau, \delta') =$$

$$= E(R(T, \delta_0) - R(T, \delta')) \geq \eta/2\, \tau(\theta_0 - \epsilon, \theta_0 + \epsilon)$$

but, since θ_0 is in the support of τ,

$\tau(\theta_0 - \epsilon, \theta_0 + \epsilon) > 0$. This contradicts the assumption that δ_0 is a Bayes rule w.r.t. τ. Hence the proof.

<u>Defn</u> 'A rule δ_0 is ϵ-admissible if there does not exist a rule δ_1 for which $R(\theta,\delta_1) < R(\theta,\delta_0) - \epsilon$ for all $\theta \in \Theta$.

<u>Exercise</u> Show that if δ_0 is ϵ-Bayes with $\epsilon > 0$ it is ϵ-admissible (no restriction on Θ or on the distribution τ w.r.t. which δ_0 is ϵ-Bayes).

Scanned with CamScanner

Definition

A set, S, in K-dimensional Euclidean space, E_K, is said to be bounded from below if there exists a finite number, M, such that for every $y = (y_1, y_2, \ldots, y_K) \in S$

$$y_j > M \quad \text{for } j = 1, 2, \ldots, K.$$

Thus a set S is bounded from below if for each fixed j, $1 \leq j \leq K$, ~~$1 \leq j \leq K$~~. y_j is bounded below as y ranges through S.

Example The set S_1 is bounded from below ~~if for each fixed j, $1 \leq j \leq K$, y_j~~

where $S_1 = \{(y_1, y_2) : y_1 y_2 = 1,\ y_2 > 0\}$

but the set $S_2 = \{(y_1, y_2) : y_1^2 = y_2,\ y_2 > 0\}$

is not bounded from below.

Proof Exercise.

Definition Let $x \in E_K$. The lower quantant at x, denoted by Q_x, is defined as the set

$$Q_x = \{y \in E_K : y_j \leq x_j, \text{ for } j = 1, 2, \ldots, K\}$$

Thus Q_x is the set of risk points as good as x and $Q_x - \{x\}$ is the set of risk points better than x.

Definition A point x is said to be a lower boundary point of a convex set $S \subseteq E_k$ if $Q_x \cap \bar{S} = \{x\}$. The set of lower boundary points of a convex set S is denoted by $\lambda(S)$.

Example The lower boundary of the unit disk $S_1 = \{(y_1, y_2) : y_1^2 + y_2^2 \le 1\}$ is

$$\lambda(S_1) = \{(y_1, y_2) : y_1^2 + y_2^2 = 1, y_1 \le 0, y_2 \le 0\}.$$

The lower boundary of the unit square.

$$S_2 = \{(y_1, y_2) : 0 \le y_1 \le 1, 0 \le y_2 \le 1\}$$

is the set consisting of ~~da~~ one point

$$\lambda(S_2) = \{(0,0)\}$$

Defn A convex set $S \subset E_k$ is said to be closed from below if $\lambda(S) \subseteq S$.

Theorem Any closed convex set is closed from below.

Pf Exercise.

Example The sets S_1 and S_2 of the above example is closed from below.

Theorem If $x = (R(\theta_1, \delta_0), \dots, R(\theta_k, \delta_0))$ is in $\lambda(S)$, then δ_0 is admissible.

Theorem If S is closed and δ_0 admissible, then $(R(\theta_1, \delta_0), \dots, R(\theta_k, \delta_0)) \in \lambda(S)$.

Exercise Give a counter example to the above theorem if S is not assumed to be closed.

<u>Theorem</u> Suppose that $\Theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$ and the risk set S is bounded from below and closed from below. Then, for every prior distribution $(p_1, p_2, \cdots, p_k)$ for which $p_j > 0$ for $j = 1, \cdots k$, a Bayes rule w.r.t $(p_1, p_2, \cdots, p_k)$ exists.

<u>Proof</u> Let $(p_1, p_2, \cdots, p_k)$ be a distribution over Θ for which $p_j > 0$ for all j and let B denote the set of all numbers of the form $b = \sum p_j y_j$ where $y = (y_1, \cdots, y_k) \in S$.

i.e., $B = \left\{ b = \sum_{j=1}^{k} p_j y_j : y = (y_1, \cdots y_k) \in S \right\}$

Since S is bounded from below, B is also bounded from below. Let b_0 be the greatest lower bound of B. In a sequence of points $y^{(n)} \in S$ for which $\sum p_j y_j^{(n)} \to b_0$, each $p_j > 0$ implies that each sequence $y_j^{(n)}$ is bounded above. Thus there exists a finite limit point y^0 of the sequence $y^{(n)}$ and $\sum p_j y_j^0 = b_0$. We now show that $y^0 \in \lambda(S)$.

Clearly, because y^0 is a limit point of points of S, $y^0 \in \bar{S}$ and $\{y^0\} \subseteq Q_{y^0} \cap \bar{S}$.

To show, $Q_{y^0} \cap \bar{S} \subseteq \{y^0\}$, if possible let $y' \in Q_{y^0} \cap \bar{S}$, $y' \neq y^0$, $\therefore \sum p_j y_j' < b_0$,

Theorem Suppose that $\Theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$ and the risk set S is bounded from below and closed from below. Then, for every prior distribution $(p_1, p_2, \cdots, p_k)$ for which $p_j > 0$ for $j = 1, \cdots k$, a Bayes rule w.r.t $(p_1, p_2, \cdots, p_k)$ exists.

Proof Let $(p_1, p_2, \cdots, p_k)$ be a distribution over Θ for which $p_j > 0$ for all j and let B denote the set of all numbers of the form $b = \sum p_j y_j$ where $y = (y_1, \cdots, y_k) \in S$.

i.e., $B = \left\{ b = \sum_{j=1}^{k} p_j y_j : y = (y_1, \cdots y_k) \in S \right\}$

Since S is bounded from below, B is also bounded from below. Let b_0 be the greatest lower bound of B. In a sequence of points $y^{(n)} \in S$ for which $\sum p_j y_j^{(n)} \to b_0$, each $p_j > 0$ implies that each sequence $y_j^{(n)}$ is bounded above. Thus there exists a finite limit point y^0 of the sequence $y^{(n)}$ and $\sum p_j y_j^0 = b_0$. We now show that $y^0 \in \lambda(S)$.

Clearly, because y^0 is a limit point of points of S, $y^0 \in \bar{S}$ and $\{y^0\} \subseteq Q_{y^0} \cap \bar{S}$.

To show, $Q_{y^0} \cap \bar{S} \subseteq \{y^0\}$, if possible let $y' \in Q_{y^0} \cap \bar{S}$, $y' \neq y^0$, $\therefore \sum p_j y_j' < b_0$, Now, since $y' \in \bar{S}$, and $\sum p_j y'_j < b_0$, there would exist points $y \in S$ for which $\sum p_j y_j < b_0$.

This contradicts the assumption that b_0 is a lower bound of B. Thus $Q_{y^0} \cap \bar{S} = \{y'\}$, implying that $y^0 \in \lambda(S)$.

Because S is closed from below, $y^0 \in S$ and the minimum value of $\sum p_j R(\theta_j, \delta)$ is achieved by a point of S. Note that any $\delta \in D^*$, for which $R(\theta_j, \delta) = y_j^0$ for $j = 1, \ldots, k$ is therefore a Bayes decision rule w.r.t. $(p_1, \ldots p_k)$.

Hence the proof.

<u>Example</u> The restriction that the prior distribution $(p_1, \ldots, p_k)$ give positive mass to each state of nature cannot be dropped.

Suppose $k = 2$ and $S = \{(y_1, y_2) : y_1 y_2 \geq 1, y_1 > 0\}$

This set is convex, bounded from below and closed from below. Consider the prior distribution (p_1, p_2) with $p_1 = 1$ and $p_2 = 0$. Then

$\sum p_j y_j = y_1$ so that the minimum Bayes risk over points of S is zero. Yet this minimum risk is not attained by a point of S, which shows that a Bayes rule w.r.t $(1, 0)$ does not exist.

Theorem

Suppose that $\Theta = \{\theta_1, \theta_2 \cdots \theta_k\}$ and that the risk set S is bounded from above and below. Also ~~at~~ S is closed from below. Then for every prior distribution $(p_1, p_2, \cdots p_k)$ (for $j=1,\cdots,k$, p_j may not be positive), a Bayes rule w.r.t $(p_1, \cdots p_k)$ exists.

Pf Exercise.

Lemma: If a nonempty convex set S is bounded from below, then $\lambda(S)$ is not empty.

Pf Exercise.

~~Exercise 1. Suppose that $\Theta = \{\theta_1, \cdots \theta_k\}$ and that the risk set S is closed from below and bounded.~~

Exercise: If S is bounded from below and closed from below and δ_0 is admissible, then $(R(\theta_1, \delta_0), \cdots, R(\theta_k, \delta_0)) \in \lambda(S)$

Exercise: Give a counter example of the above example (a) if S is not closed from below, (b) if S is not bounded from below.

Theorem Suppose that $\Theta = \{\theta_1, \dots \theta_k\}$ and the risk set S is bounded from below and closed from below. The class of decision rules
$D_0 = \{\delta \in D^* : (R(\theta_1, \delta), \dots, R(\theta_k, \delta)) \in \lambda(S)\}$
is then a minimal complete class.

Pf First we shall show that D_0 is a complete class.

Let δ be any rule not in D_0 and let
$$x = (R(\theta_1, \delta), \dots, R(\theta_k, \delta)).$$
Then $x \in S$, but $x \notin \lambda(S)$. Let $S_1 = Q_x \cap \bar{S}$; S_1 is nonempty, convex (since the closure of a convex set is convex and the intersections of two convex sets is convex), and bounded from below. Thus from previous lemma, $\lambda(S_1)$ is not empty.

Let $y \in \lambda(S_1)$; then $\{y\} = Q_y \cap \bar{S}_1$. Furthermore, $y \in Q_x$ because $y \in \bar{S}_1 = \overline{Q_x \cap S} \subseteq \bar{Q}_x = Q_x$.

Finally, $y \in \lambda(S)$ because
$$\{y\} = Q_y \cap \bar{S}_1 = Q_y \cap \overline{Q_x \cap S} = Q_y \cap Q_x \cap \bar{S}$$
$$= Q_y \cap \bar{S}.$$

Thus, because S is closed from below, there exists a rule $\delta_0 \in D_0$ for which $y = (R(\theta_1, \delta_0), \dots R(\theta_k, \delta_0))$ and which is better than δ, since $y \in Q_x - \{x\}$.

This proves D_0 complete. Therefore every rule in D_0 is admissible. Hence no proper subclass of D_0 be complete, because every complete class must contain all admissible rules. This proves D_0 minimal complete.

Definitions

A decision rule δ_0 is said to be minimax if $\sup_{\tau} r(\tau, \delta_0) = \bar{V}$, where $\bar{V}$ is the upper value defined by $\bar{V} = \inf_{\delta \in D^*} \sup_{\tau \in \Theta^*} r(\tau, \delta)$.

Similarly, a prior distribution τ_0 is said to be least favorable if

$$\inf_{\delta} r(\tau_0, \delta) = \underline{V},$$

where $\underline{V}$ is the lower value defined by

$$\underline{V} = \sup_{\tau \in \Theta^*} \inf_{\delta \in D^*} r(\tau, \delta)$$

It is always true that $\underline{V} \leq \bar{V}$.

because for all τ and all δ

$$\inf_{\delta' \in D^*} r(\tau, \delta') \leq \sup_{\tau' \in \Theta^*} r(\tau', \delta)$$

and after taking sup over $\tau \in \Theta^*$ of the left side and infimum over $\delta \in D^*$ of the right side the inequality still holds.

The Minimax Theorem

If for a given decision problem (Θ, D, R) with finite $\Theta = \{\theta_1, \cdots, \theta_K\}$ the risk set S is bounded below, then

$$\inf_{\delta \in D^*} \sup_{\tau \in \Theta^*} r(\tau, \delta) = \sup_{\tau \in \Theta^*} \inf_{\delta \in D^*} r(\tau, \delta)$$

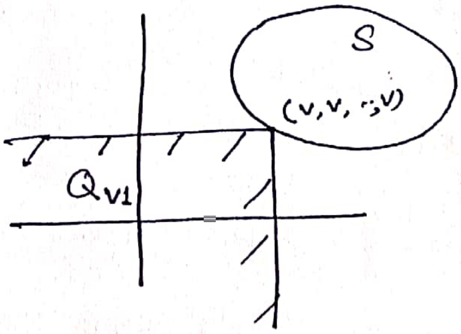
ie., $\bar{V} = \underline{V} = V$ (say),

(V is known as value of the game)

and there exists a least favorable distribution τ_0. Moreover, if S is closed from below, then there exists an admissible minimax decision rule δ_0 and δ_0 is Bayes with respect to τ_0..

Proof   It is always true that $\underline{V} \leq \bar{V}$;

Let $V = \text{lub}\{\alpha : Q_{\alpha 1} \cap S = \phi\}$ where $1 = (1, 1, \ldots, 1)^T$.



Then for every n there exists a rule δ_n s.t.

$R(\theta_j, \delta_n) \leq V + \frac{1}{n}$ for all j.

Hence $r(\tau, \delta_n) \le V + \frac{1}{n}$ for all τ and

$\sup_{\tau} r(\tau, \delta_n) \le V + \frac{1}{n}$ for all n, implying

that $\bar{V} \le V$. We will finish the proof

~~when~~ of $\bar{V} = \underline{V}$ when we show that $V \le \underline{V}$.

Denote the interior of the set Q_{V1} by Q^{0}_{V1} and note that Q^{0}_{V1} and S are disjoint convex sets, so that there must be a hyperplane $p^T x = c$, which separates Q^{0}_{V1} and S, say $p^T x \ge c$ for $x \in S$ and $p^T x \le c$ for $x \in Q^{0}_{V1}$.

Each coordinate p_j must be negative, for if $p_j < 0$ for some j we may take $x_j \to -\infty$; the other coordinates of x are fixed, so that $p^T x \to +\infty$ contradicting $x \in Q^{0}_{V1}$, $p^T x \le c$ for $x \in Q^{0}_{V1}$. Also, we may take $\sum p_j = 1$ (by dividing both sides of $p^T x = c$ by $\sum p_j > 0$). Thus p may be taken as a prior, τ_0, for nature. Because $p^T x \le c$ for all $x \in Q^{0}_{V1}$, letting $x \to V1$ implies that $V \le c$. Thus for all δ

$$r(\tau_0, \delta) = \sum p_i R(\theta_i, \delta) \ge c \ge V, \quad \text{---}(*)$$

so that $\underline{V} = \sup_{\tau} \inf_{\delta} r(\tau, \delta) \ge \inf_{\delta} r(\tau_0, \delta) \ge V$,
which complete the proof of $\underline{V} = \bar{V}$.

Again ~~Also~~ τ_0 is least favarable distribution from (*).

Now suppose, more over; that S is closed from below. Consider the δ_n as previons and let

$y_n = (R(\theta_1, \delta_n), \ldots, R(\theta_K, \delta_n))$. Because the y_n are bounded, there exists a limit point, y, of y_n. Obviously, $y \in \bar{S}$. Then $Q_y \cap \bar{S} \neq \phi$,

Nowadays since a nonempty convex set S which is bounded from must have $\lambda(S)$ not empty, we have $\lambda(Q_y \cap \bar{S}) \neq \phi$. Let $z \in \lambda(Q_y \cap \bar{S})$. Then because $Q_y \cap \bar{S} \cap Q_z = \{z\}$, we have $z \in Q_y$ and $Q_z \cap \bar{S} = \{z\}$, so that $z \in \lambda(S)$.

Because S is closed from below, $z \in S$. Any δ_0 for which $z = (R(\theta_1, \delta_0), \ldots, R(\theta_K, \delta_0))$ is admissible and satisfies $r(\tau, \delta_0) \leq V$, for $R(\theta_j, \delta_0) \leq V$ for all j.

Furthermore, $r(\tau_0, \delta_0) = V$ from (*), which shows that δ_0 is Bayes w.r.t. τ_0.

Example

Here is an example that shows if Θ is not necessarily finite $\bar{V} = \underline{V}$ of the minimax theorem does not necessarily hold.

Let $\Theta = \mathcal{A} = \{1, 2, \cdots\}$, the set of all positive integers. Let

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

~~the also~~ Assume that the random variable X is degenerate at zero for all $\theta \in \Theta$. Then

$$\sup_{\tau} r(\tau, \delta) = 1 \quad \text{for all } \delta$$

and

$$\inf_{\delta} r(\tau, \delta) = -1 \quad \text{for all } \tau.$$

So that $\underline{V} = -1 \neq \bar{V} = 1$

Show that in this example any rule $\delta \in D^*$ is minimax for the statistician.

Exercise Given an example of a game with a value (take Θ finite) for which there exists a minimax rule δ_0 that is Bayes w.r.t. some prior distribution τ_0, yet τ_0 is not least favorable. Show that if, in addition $r(\tau_0, \delta_0) = V$, then τ_0 is least favorable.

Some Theorems without proof.

Theorem 1 If δ is admissible and Θ is finite, then δ is a Bayes w.r.t. some prior distribution.

Theorem 2 If for a given decision problem (Θ, D, R) with finite Θ, the risk set S is bounded from below and closed from below, then the class of all Bayes rules is complete and admissible Bayes rules form a minimal complete class.

Theorem 3 If δ_0 is Bayes with respect to τ_0 and, for all $\theta \in \Theta$, $R(\theta, \delta_0) \leq r(\tau_0, \delta_0)$, then the game has a value, δ_0 is a minimax rule, and τ_0 is least favorable.

Theorem 4 If δ_n is Bayes w.r.t τ_n, if $r(\tau_n, \delta_n) \to C$, and if $R(\theta, \delta_0) \leq C$ for all θ, then the game has a value and δ_0 is a minimax rule.

Theorem 5 Suppose that the game (Θ, D, R) with Θ finite, has a value V and that a minimax rule δ_0 exists. Then, for any $\theta \in \Theta$ that receives positive weight from any least favorable distribution, $R(\theta, \delta_0) = V$.