Decision Theory

Classical (frequentist) statistics is based largely on the Information Theory approach: summarize the experiment with minimum loss of information.

It may be, however, that we want to use the experimental data in order to reach a decision. Then we need the Decision Theory approach.

A decision causes something to happen, something that has different consequences depending on what the true state of nature turns out to be. For example, one has to make decisions about the design of the detector, about how much time to spend on different activities, or when to publish a result. On the other hand, estimating a parameter value is not a decision, since we cannot decide what value a parameter will have.

Decision Theory gives us a rational framework for making decisions and helps us to understand the reasoning which leads to optimal decisions.

The Decision Rule

Decision theory deals with three different spaces:

An observable space χ, in which all possible observations $\mathbf{X} = (X_1, \ldots, X_N)$ fall.

A parameter space Ω contains all possible values of the parameter θ, or the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. The possible values of $\boldsymbol{\theta}$ are often called states of nature.

A decision space $\mathcal{D}$, which contains all possible decisions d.

A decision rule δ, alternatively decision procedure, decision function, specifies what decision d is to be taken given the observation $\mathbf{X}$, that is

$$d = \delta(\mathbf{X}) .$$

We shall limit ourselves to non-random decision rules δ, which define d completely when $\mathbf{X}$ is given.

Loss and Risk

To choose between decision rules one needs a loss function, $L(\theta, d)$, defined as the loss incurred in taking the decision d, when θ is the true value of the parameter (state of nature).

Alternatively, one can use the negative loss, or utility function (gain function), but it is usually more convenient and conventional to reason in terms of loss.

The expectation of the loss function $L[\theta, \delta(\mathbf{X})]$, is called the risk function for the decision rule δ:

$$R_\delta(\theta) \equiv E\{L[\theta, \delta(\mathbf{X})]\} = \int L[\theta, \delta(\mathbf{X})] f(\mathbf{X}|\theta) d\mathbf{X}\,.$$

Thus $R_\delta(\theta)$ gives the average loss over all possible observations.

Bayesian Loss and Risk

In the Bayesian framework, we can use the prior density $\pi(\theta)$ to average the risk over θ. The expected risk over θ

$$r_\pi(\delta) = \int R_\delta(\theta)\pi(\theta)d\theta$$

is called the posterior risk of using decision rule δ. This can also be written

$$\begin{aligned} r_\pi(\delta) &= E_\theta\{E_{\mathbf{X}}\{L[\theta, \delta(\mathbf{X})]|\theta\}\} \\ &= E_{\mathbf{X}}\{E_\theta\{L[\theta, \delta(\mathbf{X})]|\mathbf{X}\}\}, \end{aligned}$$

where the subscript of the expectation operator refers to the variable over which the average is taken. The quantity

$$E_\theta\{L[\theta, \delta(\mathbf{X})]|\mathbf{X}\},$$

is called the posterior loss, given the observations $\mathbf{X}$. It is the average loss over the posterior density $p(\theta|\mathbf{X})$ incurred by using the decision $\delta(\mathbf{X})$.

Example of a Decision: the Umbrella

Simple example: Decide whether to bring an umbrella on a trip.

The possible states of nature are:

- r: it rains during the trip
- $\bar{r}$: it doesn't rain during the trip

Any observables will be used to determine: $P(\text{rain})$, the (Bayesian) probability that it will rain. The loss function may be:

Decision / State of nature	Don't take umbrella	Take umbrella
no rain	0	1
rain	5	1

The obvious decision rule is to minimize the expected loss.

$$\text{Expected loss}|\text{no umbrella} = 0 \times P(\text{no rain}) + 5 \times P(\text{rain}) = 5 \times P(\text{rain})$$

$$\text{Expected loss}|\text{umbrella} = 1 \times P(\text{no rain}) + 1 \times P(\text{rain}) = 1$$

With the above loss function, the expected loss is minimized by taking an umbrella whenever the probability of rain is more than 1/5.

Example of a Decision: the Umbrella

Since the risk, or expected loss, is an average over the possible states of nature, it requires a prior probability function or prior pdf of the states of nature, so the minimum risk decision rule is a Bayesian decision rule.

There is only one important decision rule that is not Bayesian:
the minimax rule: minimize the maximum loss.

In the example above, the maximum loss $= 1$ if you take an umbrella, and is $= 5$ if you don't take it,
so the minimax decision would be always to take an umbrella.

This does not require prior probabilities, but it is a very pessimistic decision rule.

Extending the Umbrella

Usually we would have a more complicated loss function, allowing for more possible states of nature, often parameterized by some parameter θ, for example. And we could also allow for different possible decisions concerning the equipment we might carry to protect ourselves from the rain.

Such a loss function might look like:

Decision / State of nature	Don't take anything	Take Umbrella	Rubber Raincoat and Boots	Take Lifeboat
no rain	0	1	5	20
light rain	5	1	5	20
heavy rain	20	10	5	20
biblical rain	100	100	100	20

And in the limit, we could introduce the parameter $\theta =$ amount of rain, and for each decision, there would be a continuous loss function $L(\theta)$.

Choice of decision rules (frequentist)

The best decision rule is the one which gives the smallest risk $R_\delta(\theta)$.

Since the risk in general depends on the (unknown) value of the parameter θ, we can say that one decision rule is better than another only if the risk is smaller for all values of θ.

Thus if δ and δ' are two possible rules, and if

$$R_\delta(\theta) < R_{\delta'}(\theta)\,, \quad \text{for all } \theta$$

then δ is a better decision rule than δ', which we denote by $\delta > \delta'$.

In this case, the rule δ' is inadmissible.

Among admissible decision rules, the best one will depend on the (unknown) value of θ.

Bayesian decision rules

The Bayesian choice of a decision rule δ, given a prior density $\pi(\theta)$, is based on the posterior risk function

$$r_\pi(\delta) = \int R_\delta(\theta)\pi(\theta)d\theta$$

The Bayesian decision rule δ, corresponding to the prior $\pi(\theta)$, is the one which gives the smallest posterior risk, that is $r_\pi(\delta) \leq r_\pi(\delta')$ for any δ'.

Thus the Bayesian approach is to locate the uncertainty about the true value θ in a prior distribution of beliefs, $\pi(\theta)$. By averaging the risk over the posterior density, the basic classical uncertainty is avoided.

It can be shown that all admissible solutions are Bayesian solutions. Thus, if δ is an admissible decision rule, then there exists some prior density $\pi(\theta)$ such that δ is the Bayesian solution for π.

Conversely, given any decision rule, there exists some Bayesian rule which is equivalent or preferable. The class of Bayesian rules is a complete class.

Decision Rules: Minimax and Bayesian

Among non-Bayesian decision rules, the most important is the minimax method of von Neumann.

The minimax decision rule minimizes the maximum risk.

By construction, the minimax rule is admissible, if it exists. It follows that the minimax decision rule, although a classical tool, must be equivalent to a Bayesian solution for some particular prior distribution $\pi_0(\theta)$.
It can be shown that $\pi_0(\theta)$ is the most unfavourable prior distribution that can be chosen, in the sense that

$$\min\left[r_{\pi_0}(\delta)\right]_{\text{for all } \delta} \geq \min\left[r_{\pi}(\delta)\right]_{\text{for all } \delta}$$

for any prior distribution $\pi(\theta)$. In other words, the Bayesian decision rule for π_0 has a higher posterior risk than the Bayesian rule for any other prior distribution.

Thus the minimax rule leads to the most pessimistic, or conservative, decision.

Decision-theoretic Approach to Classical Problems

We have already treated estimation and testing,
essentially from the point of view of minimum loss of information
and optimal frequentist properties.

> *The alternative approach is based on decision theory. We will consider the decision theory approach to*
>
> - *point estimation*
> - *interval estimation*
> - *hypothesis testing*
> - *but not GOF testing, where decision theory doesn't apply.*

It will be seen that the decision theory approach has many interesting features, but it requires subjective input.

Point estimation by decision theory

The decision problem in point estimation is what value $\hat{\theta}$ to choose for a parameter θ, given N observations $X_1, X_2, \ldots, X_N$ from a density $f(\mathbf{X}, \theta)$.

The decision space $\mathcal{D}$ is in one-to-one correspondence with the parameter space Ω: for to each possible value θ_0, there is a decision $d : \theta = \theta_0$.

The loss is a function $L(\theta - \hat{\theta})$ of the distance between the estimate $\hat{\theta}$ and the true value θ. Clearly it must have a minimum at $\theta = \hat{\theta}$. The usual choice is a quadratic loss function of the form $L(\theta - \hat{\theta}) = \omega(\theta - \hat{\theta})^2$.

With this loss function, the posterior loss is minimal with the decision rule

$$\hat{\theta} = E(\theta|\mathbf{X}) = \int \omega(\theta - \hat{\theta})^2 f(\mathbf{X}|\theta)\pi(\theta)d\theta$$

The above is not obvious but is derived on p. 118 of the book.

For quadratic loss, the Bayesian point estimate is the mean of the posterior density of θ.

Interval estimation by decision theory

Since the posterior density $p(\theta|X)$ summarizes one's knowledge of θ, the decision problem may be to choose an interval (a, b) in the range of θ which best describes $p(\theta|X)$. A possible loss function is

$$L(\theta; a, b) = \begin{cases} \omega_1(b-a)^2 & \text{if} \quad \theta \in (a, b)\,. \\ \omega_2(\theta-a)^2 & \text{if} \quad \theta < a\,. \\ \omega_3(\theta-b)^2 & \text{if} \quad \theta > b\,. \end{cases}$$

The posterior loss now becomes

$$E[L(\theta; a, b)|X] = \omega_1 \int_a^b (b-a)^2 p(\theta|X) d\theta + \omega_2 \int_{-\infty}^a (\theta-a)^2 p(\theta|X) d\theta + \omega_3 \int_b^\infty (\theta-b)^2 p(\theta|X) d\theta\,.$$

The solution is to choose the interval (a, b) to minimize this posterior loss, which depends obviously on the values assigned to ω_1, ω_2 and ω_3.

Tests of hypotheses by decision theory

Suppose that one has to decide between two hypotheses, H_0 and H_1. Usually, one chooses a loss function which has the value 0 (no loss) for the right decision, for example:

Table 6.1. Loss function.

Decision / State of nature	Choose H_0	Choose H_1
H_0 true	0	ℓ_0
H_1 true	ℓ_1	0

Table 6.2. Probabilities of decisions, $P(H_i|\theta)$.

Decision / State of nature	Choose H_0	Choose H_1
H_0 true	$1-\alpha(\delta)$	$\alpha(\delta)$
H_1 true	$\beta(\delta)$	$1-\beta(\delta)$

Tests of hypotheses by decision theory

In the Bayesian approach, one attributes prior probabilities μ and $(1-\mu)$ to H_0 and H_1, respectively. The usual solution is then to minimize the risk

$$r_\mu(\delta) = \alpha(\delta)\ell_0\mu + \beta(\delta)\ell_1(1-\mu)\,.$$

The situation may be seen graphically by plotting the possible points $[\alpha(\delta), \beta(\delta)]$. It can be shown that the accessible region is convex, and will have the general shape as illustrated in Fig. 6.2. Bayesian decision rules correspond to points on the lower boundary of this region. For $0 < \mu < 1$, the Bayesian family is identical to the set of admissible decision rules. It is clear that the minimum risk will be obtained at the decision corresponding to the point B where the line

$$r = \ell_0\mu\alpha + \ell_1(1-\mu)\beta$$

is tangential to the region of possible points.

Hypothesis testing by decision theory

This decision is found by considering the posterior loss, given the observations **X**. Thus the expected loss in choosing H_0 is given by

$$\ell_1 P(H_1|\mathbf{X}) = \ell_1(1-\mu)P(\mathbf{X}|H_1)$$

where $P(H_1|\mathbf{X})$ is the posterior probability that H_1 is true.

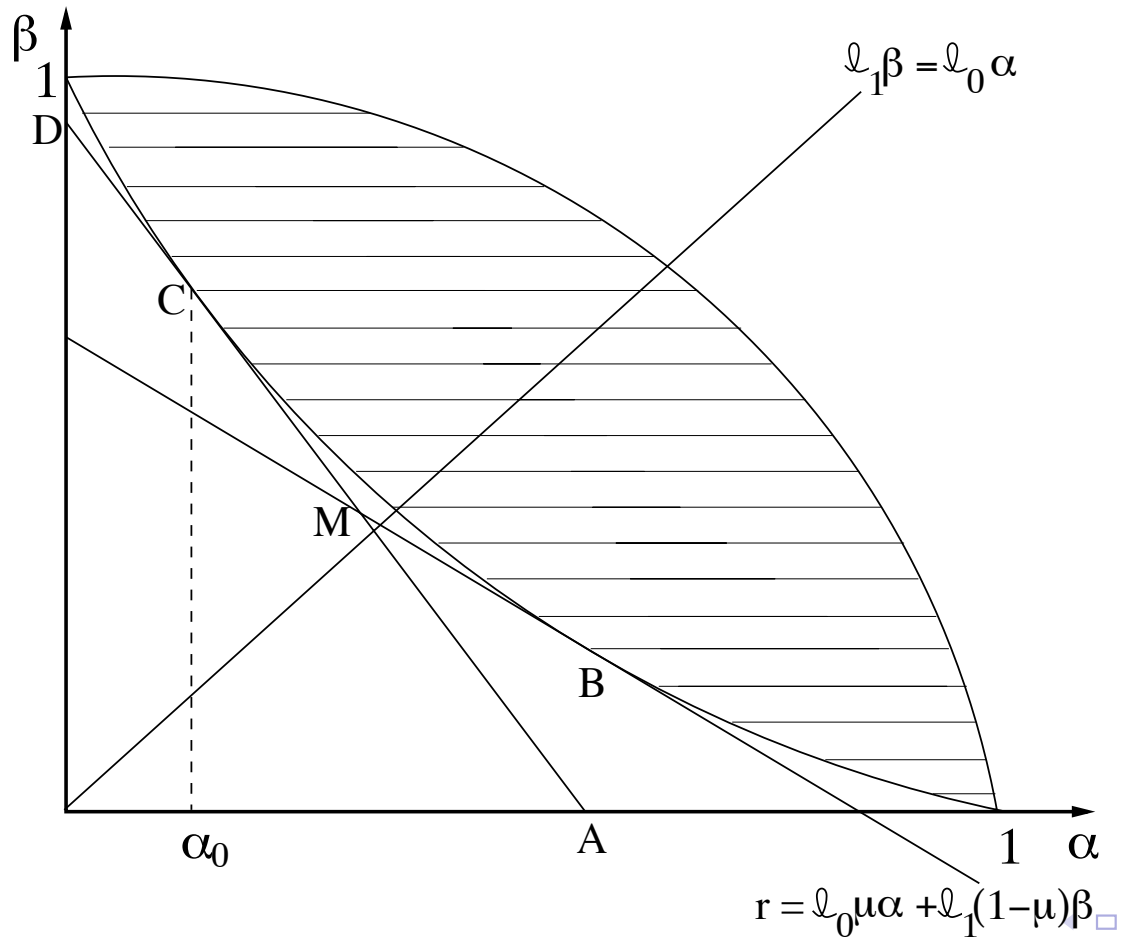

Fig. 6.2
The shaded area contains all possible points (α, β).

Hypothesis testing by decision theory

The expected loss in choosing H_1, on the other hand, is given by

$$\ell_0 \mu P(\mathbf{X}|H_0)\,.$$

Thus the minimum risk decision rule (point B in Fig. 6.2) is:

$$\text{choose } H_0 \text{ if } \quad \frac{P(\mathbf{X}|H_1)}{P(\mathbf{X}|H_0)} < \frac{\ell_0 \mu}{\ell_1(1-\mu)}\,, \qquad \text{otherwise choose } H_1$$

Classically, the decision problem would be to choose the "significance level" α_0. (see slide 11 of chap. 4)
One would then minimize β, and obtain point C. It seems obvious, and it may be shown rigorously that this procedure corresponds to a Bayesian solution with a particular choice of the ratio

$$\ell_0 \mu / \ell_1(1-\mu)\,,$$

namely, the ratio of distances OD/OA in Fig. 6.2.