Statistical Decision Theory & Inference

Statistical Experiment(s)

↓

Statistical Data → State of nature ← Decision Theory

State of nature
θ : parameter

Classical Statistics ⟶ uses only sample information collected through experiments;

but usually doesn't use extra evidence or some other extra information. But decision theory uses all kinds of informations.

Decision theory ⟶ uses prior information about θ.

Decision theory ⟶ Consequence of using/taking a particular decision

↓

Technically known as
reward / loss (in statistics)
utility (in Economics)

The basic difference between the decision theory and the classical statistics is the loss function. In decision theory we are always guided by the loss (or reward) when we estimate θ by $\hat{\theta}$. More generally, we can view the whole statistics subject from a decision theoretic point ~~of~~ justifying suitable loss functions.

Suppose $X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$, unknown.
In classical statistics, UMVUE, MLE is $\hat{\theta} = X$.
A decision theorist would ask why only $\hat{\theta} = X$?
Why not try with $\hat{\hat{\theta}}_c = cX$, $c \in (0,1)$?

When we are trying with $\hat{\theta}$ to guess θ, the error is $|\hat{\theta} - \theta|$. It may happen that for every sq.unit of deviation, we lose \$5.
i.e, total loss, L is

$$L = 5(\hat{\theta} - \theta)^2.$$

As a decision theorist, you should try to 'minimize' your loss $5(\hat{\theta}-\theta)^2$ choosing $\hat{\theta}$ suitably.

Well, we can view the use of UMVUE also from a decision theoretic point. If our loss is defined as $L = |\text{bias}| = |(E(\hat{\theta}) - \theta)|$, then there is no argument against using $X = \hat{\theta}$.

Broadly speaking, the decision theory is divided in two main schools.

Decision theory

Frequentist Approach (doesn't believe in such prior distribution)

Bayesian approach (Uses some prior distribution on θ)

Empirical Bayes approach

An example where Bayesian Approach differs from all other methods.

Keep on tossing a coin. $X = \#$ heads.
$\theta = P(H)$, $1-\theta = P(T)$, $0 \leq \theta \leq 1$.
After 12 tosses, you get 9 Heads & 3 tails.

Q. Guess θ, the probability of head.

Note: Do you really know the probability distri. of the data (9-H, 3-T)? The probability distribution depends heavily on the mechanism of the experiment.

Observe carefully, that we get 9-H, 3T because, either (i) The experiment was stopped right after 12^{th} toss or (ii) the experiment was stopped right after 3rd T.

$$(i) \implies P(9\text{-}H, 3\text{-}T) = \binom{12}{9}\theta^9(1-\theta)^3, \quad X \sim Bin(12,\theta)$$

$$(ii) \implies P(9\text{-}H, 3\text{-}T) = \binom{11}{2}\theta^9, (1-\theta)^3, \quad X \sim Neg.Bin(12,3,\theta)$$

The Bayesians treat both the models (i) & (ii) in the same way and draw same conclusion regarding θ; i.e., they are more interested in the part containing θ, which is same for both the models. Where as frequentists or the classical statisticians treat ~~both~~ those two models differently.

Using classical statistics, if we want to test $H_0: \theta = 1/2$ vs. $H_1: \theta > 1/2$

model (i) $\Longrightarrow$ ~~reject~~ Accept H_0 at 5% level

where as model (ii) $\Rightarrow$ reject H_0 at 5% level.

But if we apply (we will see) Bayesian techniques, we get the same answer from both the modes.

$X \sim NB(n, r, \theta)$: # Successes before r^{th} failure

$= P(X=x) = \binom{n-1}{r-1} \theta^x (1-\theta)^r$, $x = 0, 1, 2, \ldots\ldots$
$n = x+r$

$E(X) = r\frac{p}{q} = r\frac{\theta}{(1-\theta)}$ (ii) $n=12$, $r=3$ $E(X) = 3\frac{\theta}{(1-\theta)}$

$\Leftrightarrow 3\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) = 9$

$\Leftrightarrow \frac{\hat{\theta}}{1-\hat{\theta}} = 3$

$\Leftrightarrow \hat{\theta} = 3 - 3\hat{\theta}$

$\Leftrightarrow 4\hat{\theta} = 3$

$\Leftrightarrow \underbrace{\hat{\theta} = 3/4}$

NOT unbiased

Where as

$X \sim B(n, \theta)$

$E(X) = n\theta$ (i) $E(X) = 12\theta$

$\Rightarrow \frac{X}{12} = \hat{\theta}$

$\Leftrightarrow \underbrace{\hat{\theta} = 3/4}$

is unbiased

(i) Test $H_0: \theta = 1/2$ vs $H_1: \theta \neq 1/2$

$$\Lambda = \ln \frac{\sup_{H_0} L(\theta|X)}{\sup_{H_0 \cup H_1} L(\theta|X)} = \ln \frac{\left(\frac{1}{2}\right)^{12}}{\left(\frac{X}{n}\right)^X \left(1-\frac{X}{n}\right)^{n-X}}$$

$$\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) = \frac{\frac{X}{n}}{1-\frac{X}{n}} = \left(\frac{X}{n-X}\right)$$

If $\Lambda > c$ then accept H_0
$\leq c$ then rej H_0

i.e.,

The basic concepts of decision theory was derived from the mathematical Game theory which in the simplest situation consists of two players — Player-I and player-II. In game theory both the players are taking actions against another & the actions are called strategies. Each player has its own set of strategies.

Let, the set of strategies available to P-I be Θ, while $\mathcal{A}$ will stand for the set of possible strategies for P-II. If P-I chooses strategy $\theta \in \Theta$ and P-II chooses strategy $a \in \mathcal{A}$, there is a pay-off $L(\theta, a)$ for P-I. This is called a two-person <u>zero sum</u> game. It can be assumed that $L(\theta, a) \geq K$ (for some finite constant K)

<u>Example</u>: Two contestants simultaneously put up either one or two fingers. P-I wins if the sum of the digits showing is odd and if it is even, P-II wins. The winner receives in \$ the sum of the digits showing, this being paid by the loser.

So, $\Theta = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$

P-I Θ \ P-II $\mathcal{A}$	1	2
1	-2	3
2	3	-4

In decision theory, the role of player-I is played by the nature and P-II is the statistician. The set of all strategies for the nature is Θ, called the <u>parameter space</u> and the $\mathcal{A}$, the set of statistician's strategies is called the action space. Here, when nature chooses $\theta \in \Theta$ and the statistician chooses $a \in \mathcal{A}$, the statistician loses $L(\theta, a)$. With out loss of generality, assume $L(\theta, a) \geq 0$ $\forall a \in \mathcal{A}$, and $\theta \in \Theta$.

$$\text{So, } L : \Theta \times \mathcal{A} \longrightarrow \mathbb{R}^+.$$

Hence, we identify a statistical decision theory by the triplet $(\Theta, \mathcal{A}, L)$.

<u>Major differences between the Game theory and the decision theory.</u>

(1) The game theory, both the players are rational (assumed) and both are trying to maximize their gain (or minimizing loss) simultaneous. A common approach is the 'minimax' approach where each player is guarding against the worst situation.

$$P\text{-II} \longrightarrow \min_{a \in \mathcal{A}} \max_{\theta \in \Theta} L(\theta, a) = \overline{V}$$

$$P\text{-I} \longrightarrow \max_{\theta \in \Theta} \min_{a \in \mathcal{A}} L(\theta, a) = \underline{V}$$

Usually $\underline{V} \leq \overline{V}$

If there exists a_0 and θ_0 $\ni$

$$\max_{\theta \in \Theta} L(\theta, a_0) = \min_{a} \max_{\theta} L(\theta, a)$$

$$= \max_{\theta} \min_{a} L(\theta, a) = \min_{a} L(\theta_0, a)$$

i.e., $\underline{V} = \overline{V} \underset{(\text{say})}{=} V$ we say that the game is determined and the value of the game is V.

In decision theory, the nature is not 'rational'. It chooses $\theta \in \Theta$ arbitrarily, or it doesn't plan to maximize $L(\theta, a)$. Its the statistician's turn to minimize $L(\theta, a)$. In this sense, it is a 'kind of' one sided game (though the minimax principle can be applied).

(2) There is another constraint for the statistician. Before, taking any decision, he performs a statistical experiment whose outcome is random, X and depending on X, he chooses an optimal action.

<u>Moreovere</u>, ~~the~~ ~~di~~ the out come X has a probability distribution P_θ, $\theta \in \Theta$ which depends on nature's choice θ ; i.e., the

gets some information about the state of the nature. So, the statistical decision theory, strictly speaking, is not a fair game.

Let $\mathfrak{X}$ be the sample space i.e., the outcome X of the stat. exp. takes values in $\mathfrak{X}$. Instead of using a, to denote the action of the statistician, we use $\delta(X)$ which depends on θ. So our loss function is

$$L(\delta(X), \theta) \geq 0$$

and let $\mathcal{D} = \{$ set of all possible δ's $\}$.

$\delta(X)$ is called a '<u>rule</u>' (decision rule).

<u>Basic assumption</u>: $L(\delta(X), \theta)$ is convex in $(\delta(X) - \theta)$.

Our goal is to choose $\delta(X)$ such that $L(\delta(X), \theta)$ is "small". But $L(\delta(X), \theta)$ depends on X, so the loss is again a "random quantity". So, we try to minimize average loss, called risk, defined as

$$R(\delta, \theta) = E_\theta\, L(\delta(X), \theta).$$

So, to minimize $R(\delta, \theta)$, it's not a question of choosing a particular value $\delta(X)$, it's the question of choosing the structure or functional form δ of the decision rule.

Loss functions ; (Two general Loss functions)

1. Squared error loss (quadratic loss)

The loss function ~~L(θ, δ)~~ = ~~(θ - a)~~

$$L(\delta(x), \theta) = (\delta(x) - \theta)^2$$ is called

squared error loss.

Why should we use squared error loss ? (SEL)

(i) SEL was first introduced in estimation problems when unbiased estimators of θ being considered, since, $R(\delta, \theta) = E(\delta(x) - \theta)^2$ would then be the variance ; otherwise, its simply MSE.

(ii) Another reason for the popularity of SEL is due to the relationship to classical least squares theory which in turn connects the maximum likelihood estimation method under normality

(iii) Finally, the SEL makes the calculations much easier.

In the multivariate setup when we estimate $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_p)'$, by $\underset{\sim}{\delta}(\underset{\sim}{X}) = (\delta_1(\underset{\sim}{X}), \ldots, \delta_p(\underset{\sim}{X}))$, the natural generalization is

$$L(\underset{\sim}{\delta}(\underset{\sim}{X}), \underset{\sim}{\theta}) = \| \underset{\sim}{\delta}(\underset{\sim}{X}) - \underset{\sim}{\theta} \|^2$$
$$= (\underset{\sim}{\delta}(\underset{\sim}{X}) - \underset{\sim}{\theta})' (\underset{\sim}{\delta}(\underset{\sim}{X}) - \underset{\sim}{\theta}).$$

(iv) Another justification for the squared error loss is :—
Let the loss be a fn. of the deviation ; i.e.,

$$L(\delta, \theta) = g(\delta - \theta) \quad (\text{by Taylor's expansion series})$$

$$= g(0) + (\delta-\theta)\, g'(0) + \tfrac{1}{2}(\delta-\theta)^2\, g''(0)$$

(neglecting the higher order terms of $(\delta-\theta)$)

$$\Longrightarrow L(\delta, \theta) \approx K_1\left((\delta-\theta) + K_2\right)^2 + K_3$$

for suitable constants K_1, K_2, K_3

So, basically $L(\delta,\theta) \propto (\delta + c - \theta)^2$

if we redefine $\delta^* = \delta + c$

then minimizing $L(\delta,\theta)$ w.r.t. δ is same as minimizing $L^*(\delta^*, \theta) = (\delta^* - \theta)^2$ which brings us to SEL again.

(See Page - 60 , Berger)

---

2. <u>0-1 Loss</u>

In 0-1 loss, the decision space $\mathcal{D}$ consists of two elements ; i.e., $\mathcal{D} = \{\delta_0, \delta_1\}$. We want to guess θ, $\theta \in \Theta$. More over $\Theta = \Theta_0 \cup \Theta_1$ (disjoint)

The loss is described as follows —

$$\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$$

we take δ_0 if $X \in \mathcal{X}_0$ & δ_1 if $X \in \mathcal{X}_1$.

$$L(\delta_0, \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ 1 & \text{if } \theta \in \Theta_1 \end{cases}$$

$$L(\delta_1, \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_1 \\ 1 & \text{if } \theta \in \Theta_0 \end{cases}$$

So, the risk fn. is given as

$$R(\delta_0, \theta) = E\, L(\delta_0, \theta) = P(X \in \mathcal{X}_0 \mid \theta \in \Theta_1)$$
$$= \text{Type-II error} = 1 - \text{Power}.$$

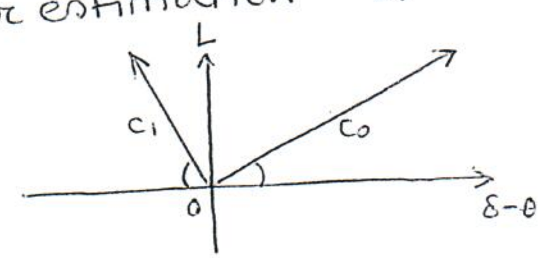
$$R(\delta_1, \theta) = E\, L(\delta_1, \theta) = P(X \in \mathcal{X}_1 \mid \theta \in \Theta_0)$$
$$= \text{Type-I error} = \text{size/level}.$$

This is infact the classical hypothesis testing. (See Ferguson, 198)

Apart from the squared error loss, in point estimation, another common loss is 'Linear loss' given as —

$$L(\delta, \theta) = \begin{cases} c_0\, (\delta - \theta) & \text{if } \delta \geq \theta \\ c_1\, (\theta - \delta) & \text{if } \delta < \theta \end{cases}$$

The constants c_0 & c_1 can be chosen to reflect the relative importance of overestimation or underestimation.



If $c_0 = c_1 = c$,

it is absolute error loss $\propto |\delta - \theta|$.

There are several other loss functions, like entropy losses, invariant squared error losses etc. which will be discussed when necessary.

Criticisms: The use of loss functions is often criticised for

(i) being inappropriate for the inference problem

(ii) Even if one derives a loss function from the economic utility function, it may be extremely difficult to handle.

(iii) Loss functions are 'non robust', in the sense that, ~~an~~ a "good" estimator for the loss L_1 may not be good for another loss L_2.

To tackle (i), we should concentrate on most intuitive loss functions; however we have to admit our limitations regarding (ii). But we can do something for (iii). We can look for estimators which perform more or less "well" under various loss functions and are called robust estimators — a major field of research.

Admissibility & Completeness

Def 1. (a) A decision rule δ_1 is said to be as good as rule δ_2 if $R(\delta_1, \theta) \le R(\delta_2, \theta)$ $\forall \; \theta \in \Theta$.

(b) A rule δ_1 is said to be better than a rule δ_2 if $R(\delta_1, \theta) \le R(\delta_2, \theta)$ $\forall \; \theta \in \Theta$ and $R(\delta_1, \theta) < R(\delta_2, \theta)$ for some θ.

(Notation: $\delta_1 \succ \delta_2$)

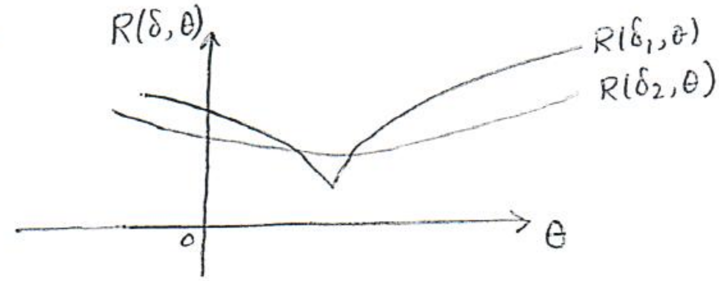
(c) A rule δ_1 is said to be equivalent to a rule δ_2 if $R(\delta_1, \theta) = R(\delta_2, \theta)$ $\forall \; \theta \in \Theta$.

A above definitions give a partial ordering of the space $\mathcal{D}$ of decision rules.

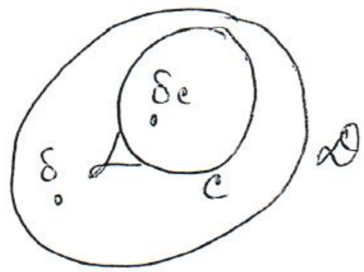
Def - 2. A rule δ is said to be admissible if there exists no rule better than δ.

A rule is said to be inadmissible if it is not admissible.

Admissibility is an optimum property, although in a very weak sense. Conversely, we could never feel very proud about a rule if it is inadmissible.



Def-3. (a) A class $\mathcal{C}$ of decision rules, $\mathcal{C} \subset \mathcal{D}$, is said to be complete, if, given any $\delta \in \mathcal{D}$, $\delta \notin \mathcal{C}$, $\exists$ a rule $\delta_c \in \mathcal{C}$ $\ni$ δ_c is better than δ.



(b) A class $\mathcal{C}_*$ of decision rules is said to be essentially complete, if, given any rule $\delta \in \mathcal{D}$, $\delta \notin \mathcal{C}$, $\exists$ a rule $\delta_* \in \mathcal{C}_*$ $\ni$ δ_* is as good as δ

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

Pf:- (Ex)

Lemma 2. If $\mathcal{C}_*$ is an essentially complete class and there exists an admissible rule $\delta \notin \mathcal{C}_*$, then $\exists$ $\delta' \in \mathcal{C}_*$ which is equivalent to δ.

Pf:- (Ex)

Def-4. A class $\mathcal{C}_0$ of decision rules is said to be minimal complete if $\mathcal{C}_0$ is complete and if no proper subclass of $\mathcal{C}$ is complete; i.e., $\mathcal{C}_0$ is the smallest complete class.

Similarly we can define minimal essentially comple class.

Theorem 1. If a minimal comple class exists, it consists of exactly the admissible rules.

Pf:- (See Ferguson, p-56).

Minimaxity

Def 5. (a) A decision rule δ_0 is said to be minimax if
$$\sup_{\theta \in \Theta} R(\delta_0, \theta) = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\delta, \theta) .$$

The value on the right hand side is called the minimax value.

(b) Let $\epsilon > 0$. A decision rule δ_0 is said to be ϵ - minimax if
$$\sup_{\theta \in \Theta} R(\delta_0, \theta) \leq \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\delta, \theta) + \epsilon$$

Geometric interpretation:

First we define the risk set S
$$S = \{ (\ldots, R(\delta, \theta), \ldots) \; \theta \in \Theta \mid \delta \in \mathcal{D} \}$$

If Θ is finite, say $\Theta = \{\theta_1, \theta_2\}$

then, $S = \{ (R(\delta, \theta_1), R(\delta, \theta_2)) \mid \delta \in \mathcal{D} \}$

for rule δ, $\sup_{\theta \in \Theta} R(\delta, \theta) = R(\delta, \theta_1) \vee R(\delta, \theta_2)$