U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 301N /MASTAT – 301N

Decision Theory & Bayesian Analysis

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Pramendra Singh Pundir **Writer**
Department of Statistics
University of Allahabad, Prayagraj

Prof. G. S. Pandey (Rtd.**)** **Editor**
Department of Statistics
University of Allahabad, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 301N/ MASTAT – 301N **DECISION THEORY & BAYESIAN ANALYSIS**
©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Blocks & Units Introduction

The present SLM on ***Decision Theory and Bayesian Analysis*** consists of eleven units with three blocks.

The ***Unit - 1 – Introduction to Decision Theory & Bayesian Analysis,*** is the first unit of present self-learning material, which describes some basic concepts, along with their importance and scope with suitable examples.

The ***Block - 1 – Basic Elements and Bayes Rules,*** is the first block, which is divided into three units, and deals with the fundamentals of decision theory.

In ***Unit – 2 – Basic Elements***, is mainly emphasising on the basic elements of decision theory in order to create a conceptual clarity.

In ***Unit – 3 – Bayes and Minimax Rules***, focuses mainly on a comparative study of Bayes and minimax rules, with a goal to make the real-world usefulness of these rules clear to learners.

In ***Unit – 4 – Bayesian Interval Estimation,*** is being introduced the interval estimation from Bayesian perspective. Also, this unit compares the same with the classical approach.

The ***Block - 2 – Optimality of Decision Rules*** is the second block with four units, and focuses on equipping the learner with the knowledge about the optimality criteria for decision rules in Bayesian framework.

In ***Unit – 5 – Admissibility and Completeness***, discusses the concept and criteria for admissibility and completeness of decision rules. The object of this exercise is to give the learner a sight to ensure the goodness of decisions.

In ***Unit – 6 – Minimaxity and Multiple decision Problem*** has been introducing the problem of minimaxity, and the problem of making a decisions out of different available options.

Unit – 7 – Bayesian Decision Theory explores the decision theory in a Bayesian manner. So this unit discusses different aspects from a Bayesian perspective.

Unit – 8 – Bayesian Inference dealt with the problem of inference in Bayesian Scenario.

The ***Block - 3 – Bayesian Analysis*** has three units. This block comprises

Unit – 9 – Prior and Posterior Distributions, focuses on giving an insight about the prior and posterior distribution to the learner. After this one will find oneself ready to choose a suitable prior necessary for performing the Bayesian analysis.

In ***Unit – 10 – Bayesian Inference Procedures***, discussed the inferential procedures in addition to Unit-8 of Block-2.

Unit – 11 – Bayesian Robustness, discussed the concept of Bayesian robustness and focuses on explaining how this concept helps the Bayesians to ensure the firmness of their decisions. Furthermore, this unit discusses the MCMC methods for Bayesian calculations.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT – 1: INTRODUCTION TO DECISION THEORY & BAYESIAN ANALYSIS

Structure

1.1 Introduction

The world is full of uncertainty. And making a good decision in this uncertainty has always been a challenge for the humanity. This Unit discusses about a few most popular and broader classes of decision policies and their basis.

1.2 Objectives

After studying this unit, you should be able to

- Explain types of decisions.
- Classify the decision problems from the perspective of a statistician.
- Define various decision policies of importance.
- Describe Bayesian criteria for decision making.

1.3 Various Aspects of Decision Making

Consider an example where the game being played only has a maximum of two possible moves per player each turn. Then, obvious policy of a player will be of maximizing the benefits, and the moves of the opponent will aim to minimize the gains of the first player. Thus, the decision-making process takes into account all the possible observations or information. And hence it involves the making of a decision to a categorical proposition, intended to achieve particular goals.

The **optimistic approach** would be the one that evaluates each decision alternative in terms of the best payoff that can occur. The decision alternative that is recommended is the one that provides the best possible payoff. For a problem in which maximum profit is desired, the optimistic approach would lead the decision maker to choose the alternative corresponding to the largest profit. For problems involving minimization, this approach leads to choosing the alternative with the smallest payoff. Similarly, the **conservative approach** evaluates each decision alternative in terms of the worst payoff that can occur. The decision alternative recommended is the one that provides the best of the worst possible payoffs. For a problem in which the output measure is profit, the conservative approach would lead the decision maker to choose the alternative that maximizes the minimum possible profit that could be obtained. For problems involving minimization, this approach identifies the alternative that will minimize the maximum payoff. Another one is, **minimax regret approach** to decision making where one would choose the decision alternative that minimizes the maximum state of regret that could occur over all possible states of nature. This approach is **neither purely optimistic nor purely conservative**.

In statistics we refer to another approach, based on prior information, and observations as well as the assessment of the risk associated with each decision, called the **Bayesian Decision makin**g. This approach makes use of the famous Bayes theorem.

1.4 Bayes theorem and Bayesian Statistics

Bayes' theorem is named after the Reverend Thomas Bayes, a statistician and philosopher of 18th century. Bayes used conditional probability to provide an algorithm that uses evidence to calculate limits on an unknown parameter. For any two disjoint events A and B, the Bayes' theorem is stated mathematically as: $P(A \mid B)= P(B \mid A)P(A)/P(B)$. (Proof can be seen from any graduate level text). Thus, this theorem enables the user to move backward in the light of presently available observations and the prior information about the unknown parameter. The whole theory of Bayesian statistics is based on this fundamental theorem. **Bayesian statistics** is a theory in statistics based on the Bayesian interpretation of probability i.e. probability expresses some degree of belief in an event. This degree of belief may be based on prior knowledge about the event, obtained as the results of previous experiments, or on personal beliefs (called subjectivity) about the event.

1.5 Self- Assessment Exercise

1. Discuss about various real world situations and decision policies used by the decision makers.
2. State Bayes theorem and explain how does it help in decision making.

1.6 Summary

In our day to day life we come across a number of decision making situations. And there we take a decision that suits most to our objectives. Different situations and logics affect our decisions. In section 1.3, some of the most popular situations have been discussed. Section 1.4 explains the basis of such a policy in Bayesian sense followed by a few exercises, summary of the unit and a list of suggested readings.

1.7 Further Reading

1. Berger, J.O. (1985). Statistical decision theory-Fundamental concepts and methods, Springer Verlag.
2. Ferguson, T.S. (1967). Mathematical statistics- A decision theoretic approach, Academic press.
3. Lindley, D.V. (1965). Introduction to probability and statistical inference from Bayesian view point, Cambridge university press.

U.P.RajarshiTandon Open University, Prayagraj

MScSTAT – 301N /MASTAT – 301N Decision Theory & Bayesian Analysis

Block: 1 Basic Elements and Bayes Rules

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. RajarshiTandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P.RajarshiTandon Open University, Prayagraj

Course Preparation Committee

Dr. Pramendra Singh Pundir **Writer**
Department of Statistics
University of Allahabad, Prayagraj

Prof. G. S. Pandey (Rtd.) **Editor**
Department of Statistics
University of Allahabad, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. RajarshiTandon Open University, Prayagraj

MScSTAT – 301N/ MASTAT – 301N DECISION THEORY & BAYESIAN ANALYSIS

©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Block & Unit Introduction

The present block of this SLM consists of three units.

The ***Block - 1 –Basic Elements and Bayes Rules,*** is the first block, which is divided into three units,

In ***Unit – 2 –Basic Elements***, the main emphasis is given to the basic elements of Bayesian theory

Unit – 3 –Bayes and Minimax Rules, is focusing mainly on these rules.

In ***Unit – 4 – Bayesian Interval Estimation,*** is being introduced the interval estimation in Bayesian context.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT-2: BASIC ELEMENTS

Structure

2.1 Introduction

Decision theory is the study of the reasoning underlying any decision. Statistical Decision theory may be considered as the theory of making decisions in the presence of statistical knowledge. In section 2.3, we shall consider a game problem to make the decision theoretic problem and related concepts clear.

2.2 Objectives

After studying this unit, you should be able to

- Explain decision problem as a game problem.
- Explain the decision problem from the perspective of a statistician.
- Define various components and topics of importance.
- Describe Bayes and minimax criteria.

2.3 Decision Theoretic Problem as a Game Problem and Basic Elements

Suppose, you want to buy a new mobile phone. How do you decide which one is best for you and from where to buy it? That is a decision problem. Now suppose that you have, anyhow finalized the mobile you are willing to have. Then, **Decision theory** is the study of the reasoning underlying this decision. It is closely related to the well-known theory of games. In this chapter, firstly a decision problem has been explained as a game problem. Then it is explained from the perspective of a statistician. Various elements/components along with some other topics of importance have also been defined in this section. Next this chapter is focused on Bayes and minimax criteria and their description.

2.3.1 Game Theory and Decision Theory:

Basic Elements: the elements of decision theory are similar to those of the theory of games. In particular, decision theory may be considered as the theory of two-person game, in which nature takes the role of one of the players. The so-called normal form of a zero-sum two-person game, henceforth to be referred to as a *game,* consists of three basic elements:

1. A non empty set, Θ, of possible states of nature, sometimes referred to as the parameter space.
2. A non-empty set, a, of action available to the statistician.
3. A loss function, L (θ, a), a real-valued function defined on Θ X a.

A game in mathematical sense is just such a triplet (Θ, a, L), and any such triplet defines a game, which is interpreted as follows.

Nature choose a point θ in Θ, and the statistician, without being informed of the choice nature has made, chooses an action a in a. as a consequence of these two choices, the statistician loses an amount L (θ, a).[the function L (θ, a) may take negative values. A negative loss may be interpreted as a gain, but throughout this book L (θ, a) represented the loss to the statistician if he takes action a when θ is the ‘’ *true state of nature*’’.] Simple through this definition may be, its scope is quite broad, as the following example illustrated.

Example2.1: **Odd or even:** two contestants simultaneously put up either one or two fingers. One of the players, call him 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 (Θ, a, L), out of this game we give player I the label ''nature'' and the player II the label ''statistician''. Each of these players has two possible choices, so that Θ= {1, 2} =a, in which ''1'' and ''2'' stands for the decision to put up one and two fingers, respectively. The loss function is given by the table 1.1.

Thus L (1, 1) =-2

Table 2.1

	a	1	2
Θ	1	-2	3
	2	3	-4

L (1, 2) =3, L (2, 1) =3 and L (2, 2) =-4 it is quite clear that this is a game in the sense described in the first paragraph. This example is discussed later, in which it is shown that one of the players has a distinct advantage over the other. Can you tell which one it is? Which player would you rather be?

Example2.2: Consider the game (Θ, a, L) in which$\Theta = (\theta_1, \theta_2)$, a =(a_1, a_2) and the loss function L is given by the table 1.2:

(Table 2.2)

'Statistician'

		a_1	a_2
'nature'	θ_1	4	1
	θ_2	-3	0

In game theory, in which the player choosing a point from Θ is assumed to me intelligent and his winnings in the game are given by the function L (loss function of the statistician or gain function of the nature), 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 chooses θ_2 .thus it is clear that the statistician should choose action a_2 instead of action a_1 , for he will lose only one instead of four. This is the only reasonable things for him to do.

Now, suppose that the function L does not reflect the winning of nature or that nature chooses a state without any clear objective in mind. Then we can no longer state categorically that the statistician should choose action a_2 if nature happens to chooses θ_2 , the statistician will prefer take action a_1.

2.3.2 Decision Function & Risk Function:

To give a mathematical structure to this process of information gathering, we suppose that 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. The sample space denoted as $\mathfrak{X}$ 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 Borel subsets, β of $\mathfrak{X}$. thus for each θ ϵ Θ there is a probability measure P_θ defined on β, a corresponding cumulative distribution function $F_X(x/\theta)$ which represents the distribution function of X when θ is the true state of the nature (the parameter)

A statistical decision problem or a statistical game is a game (Θ, a, L) coupled with an experiment involving a random variable X whose distribution P_θ depends on the state θЄΘ chosen by nature.

On the basis of the outcome of the experiment X=x (x is the observed value of X), the statistician chooses an action d(x)ϵ a .such a function d, which maps the sample space $\mathfrak{X}$ in to a, is an elementary strategy for the statistician in this situation .The loss is now the random quantity L (θ, d(x)).The expected value of L (θ, d(x)) when θ is the true state of nature is called the risk function.

$$R(\theta, d) = E\{L(\theta, d(x))\} \text{.........................} \quad (2.1)$$

and represented the average loss to the statistician when the true state of nature θ and the statistician used the function d.

<u>Defn.</u> 2.1: Any function d(x) that maps the sample space $\mathfrak{X}$ in to a, is called a non-randomized decision rule or a non-randomized decision function, provided the risk function R (θ, d) exists and is finite for all θϵΘ. The class of all non-randomized decision rules is denoted by D.

$$R(\theta, d) = E_\theta L(\theta, d(x)) = \int L(\theta, d(x)) dP_\theta\,(x) \text{..............} \quad (2.2)$$

With such an understanding, D consists of those functions d for which $L(\theta, d(x))$ is for each θϵΘ a Lebesgue integrable function of x. In particular, D contains all simple functions. On the other hand, the expectation in (2.2) may be taken as the Riemann or the Riemann-Stieltjes integral.

$$R(\theta, d) = E_\theta L(\theta, d(x)) = \int L\left(\theta, d(x_j)\right) dF_x\,(x/\theta) \text{........} \quad (2.2)$$

In that case D would contain only functions d for which $L(\theta, d(x))$ is for each θϵΘ continuous on a set of probability one under $F_x(x/\theta)$.

<u>Example</u> 2.1: 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 the value 1 or 2. If θ=1 is the true state of nature, $P_{\theta=1}^{[X=1]} = \frac{3}{4} = 1 - P_{\theta=1}^{[X=2]}$. Similarly $P_{\theta=2}^{[X=1]} = 1/4 = 1 - P_{\theta=2}^{[X=2]}$. There are exactly four possible functions from $\mathfrak{X}$={1,2} in to, a={1,2}. There are the four decision rules,

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

Rules $d_1$and$d_4$ignore the value of X , ruled_2 reflects the belief of the statistician that the nature is telling the truth, and ruled_3, that nature is not telling the truth. The risk Table (2.1) is given as:

(Table 2.1)

D

		d_1	d_2	d_3	d_4	
	1	-2	$-3/4$	$7/4$	3	
Θ	2	3	$-9/4$	$5/4$	-4	←R(θ,d)

It is a custom, which we steadfastly observe, that the choice of a decision function should depend only on the risk function $R(\theta, d)$ and not other wise on the distribution of the random variable$L\big(\theta, d(X)\big)$.

Notice that the original game (Θ, 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 a, L, and the distribution of X, whose expectation must be the main objective of decision theory.

A ''classical'' mathematical statistics consists three important categories:

1. *a Consists of two points*,$a = \{a_1, a_2\}$: decision theoretic problems in which a consists of exactly two points are called *problem of testing hypothesis*.

Consider the special case in which Θ is the real line and suppose that the loss function for some fixed number θ_0 given by the formulas:

$$L(\theta, a_1) = \begin{cases} l_1 if \theta > \theta_0 \\ 0 \quad if \theta \le \theta_0 \end{cases} \text{and} L(\theta, a_2) = \begin{cases} 0 \quad if \theta > \theta_0 \\ l_2 if \theta \le \theta_0 \end{cases}$$

Where l_1 and l_2 are positive numbers. Here we would like to take action $a_1\ if \theta \le \theta_0$ and action $a_2\ if \theta > \theta_0$.the space D of decision rule consists of those functions d from the sample space in $\{a_1, a_2\}$ with the property that $P_\theta[d(x) = a_1]$ is well-defined for all values of θϵΘ. The risk function in this case is ,

$$R(\theta, d) = EL\big(\theta, d(x)\big)$$

$$= l_1 P_\theta[d(x) = a_1\ ] if \theta > \theta_0$$

$$= l_2 P_\theta[d(x) = a_2\ ] if \theta \leq \theta_0$$

In this case probabilities of making two types of error are involved. For $\theta > \theta_0$,$P_\theta[d(x) = a_1\ ]$ is the probability of making the error of taking action a_1 when we should take action a_2 and θ is the true state of nature. Similarly, for $\leq \theta_0 P_\theta[d(x) = a_2\ ] = 1 - P_\theta[d(x) = a_1\ ]$, is the probability of making the error of taking action a_2 when we should take action a_1 and θ is the true state of nature.

2. *a Consists of k points,*$\{a_1\ , a_2\ , \ldots \ldots a_k\}, k \geq 3.$ these decision theoretic problems are called *multiple decision problems*. For an example an experimenter is to judge which of treatments has a greater yield on the basis of an experiment.

He may (a) decide treatment 1 is better, (b) decide treatment 2 is better, or (c) withhold judgment until more data are available. In this exp. k=3

3. *a Consists of a real line,* $a = (-\infty, \infty)$.

such decision theoretic problems are referred to in a board sense as *point estimation of a real parameter*. Consider the special case in which Θ is also a real line and suppose that the loss function is given by the formula,

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

Where, c is some positive constant. A decision function d, in this case a real–valued function defined on a 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.

$$R(\theta, d) = EL\big(\theta, d(x)\big)$$

$$= cE_\theta\big(\theta - d(x)\big)^2 ,$$

The criterion arrived here is that of choosing an estimate with a small mean squared error in some sense.

2.3.3 Randomization:

It is often useful to recognize explicitly that in any decision problem, the statistician may wish to choose a decision from D by means of an auxiliary randomization procedure of some short, such as by tossing a coin. In other words, the statistician may wish to make a mixed or randomized decision δ by assigning probabilities $p_1, p_2, \dots \dots$ to the elements $d_1, d_2, \dots \dots$ of decisions from D and then one of the decisions δ on the basis of these probabilities is chosen.

More generally, a randomized decision for the statistician in a game (Θ, a, L) is a probability distribution over a (it is understood that a fixed σ-field of subsets of $a$containing the individual points of $a$is given). If P is probability distribution over a and Z is a random variable taking values is a.whose distribution is given by P, the expected or average loss in the use of randomized decision P is,

$$L(\theta, P) = EL(\theta, Z) \dots\dots\dots\dots\dots\dots \quad (3.1)$$

Provided it exists. This formula is to be regarded as an extension of the domain of definition of the function $L(\theta, \cdot)$ from$a$ to the sample space of randomized decisions, for each element a ϵa may, and shall, be regarded as the probability distribution degenerate at a ,that is, the distribution giving probability one to point a. *the space of randomized decisions, P, for which$L(\theta, P)$ exists and is finite for all θϵΘ* is *denoted by a^** .

With this definition, the game (Θ,a^*, L) is to be considered as the game (Θ, a, L) in which the statistician is allowed randomization. a^**c*ontains all the probability distributions giving mass one to a finite number of points of$a$.

By analogy, we may extend the game (Θ, D, R) to (Θ, D^*, R) whereD^* is a space containing probability distribution over D. if δ denotes a probability distribution over D, R(θ ,δ) is defined analogously to (3.1) as,

$$\mathrm{R}(\theta, \delta) = \mathrm{E}\,\mathrm{R}(\theta, \mathrm{Z}) \dots\dots\dots\dots\dots \quad (3.2)$$

Where Z is a random variable taking values in D, whose distribution is given by δ.

Defn.3.1: Any probability distribution δ on the space of non-randomized function, D, is called a randomized decision function or a randomized decision rule, provided the risk function (3.2) exists and is finite for all θϵΘ. The space of all randomized decision rule is denoted by D*. D* contains all the probability distributions giving mass one to a finite number of point of D.

The space D of non-randomized decision rules may, and shall, be considered as a subset of the space D* of randomized decision rules D ϵ D* by identifying a point d ϵ D with the probability distribution δ ϵ D* degenerate at point d.

One advantage in the extension of the definition of L (θ, ·) from $a to a^*$ and the definition of R (θ, ·) from D to D* is that these functions become linear on $a^*$and D*, respectively. In other words, if$P_1 \epsilon a^*$,$P_2 \epsilon a^*$ and $0 \leq \alpha \leq 1$.

$$P = \alpha P_1 + (1-\alpha)P_2 \epsilon a^* \text{and} L(\theta, \alpha P_1 + \overline{1-\alpha} P_2) = L(\theta, P) = EL(\theta, Z)$$

$$= \alpha L(\theta, P_1) + (1-\alpha)L(\theta, P_2) \ldots\ldots\ldots\ldots\ldots (3.3)$$

Similarly, if$\delta_1 \epsilon a^*$,$\delta_2 \epsilon a^*$ and $0 \leq \alpha \leq 1$.then

$$\delta = \alpha\delta_1 + (1-\alpha)\delta_2 \epsilon D^*$$

$$R(\theta, \delta) = ER(\theta, Z) = \alpha R(\theta, \delta_1) + (1-\alpha)R(\theta, \delta_2) \ldots\ldots\ldots\ldots\ldots\ldots (3.4)$$

Example3.1: Let the game be defined as,

		a_1	a_2	a_3
Θ	θ_1	4	1	3
	θ_2	1	4	3

If nature choosesθ_1, actiona_3 is preferable to action a_1.if, on the other hand, nature choosesθ_2, actiona_3 is preferable to action $a_{2.}$ thusa_3 is preferred to either of the other action under the proper circumstances. However, suppose the statistician flips a fair coin to choose between actions a_1 anda_2; that is suppose the statistician's decision is to choose $a_1$if the coin comes up heads and choose $a_2$if the coin comes up tails. This decision, denoted by δ, is a *randomized* *decision*; such decisions allow the actual choice of the action in 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 chooses action a_1 with probability ½, action a_2 with probability ½, actiona_3 with probability zero. The expected loss in the use of δ is given by,

$$L(\theta, P) = EL(\theta, Z) = 1/2L(\theta, a_1) + 1/2L(\theta, a_2) + 0L(\theta, a_3)$$

$$= \frac{1}{2}.4 + \frac{1}{2}.1 + 0.3 = \frac{5}{2} if \theta = \theta_1$$

$$= \frac{1}{2} + \frac{4}{2}.1 + 0.3 = \frac{5}{2} if \theta = \theta_2$$

Because it is understood that the choice between strategies is to be made on the basis of expected loss only, δ is certainly to be preferred toa_3 for no matter what the true state of nature, the expected loss is smaller if we use δ than if we usea_3.

$$P_1 = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right), \qquad P_2 = \left(\frac{3}{8}, \frac{5}{8}, 0\right)$$

$$L(\theta, P_1) = \frac{4}{4} + \frac{1}{2} + \frac{3}{4} = \frac{9}{4} if \quad \theta = \theta_1$$

$$= \frac{1}{4} + \frac{4}{2} + \frac{3}{4} = \frac{12}{4} if \theta = \theta_2$$

$$L(\theta, P_2) = \frac{3}{8}.4 + \frac{5}{8}.1 + 0.3 = \frac{17}{8} if \quad \theta = \theta_1$$

$$= \frac{3}{8} + \frac{5}{8}.4 + 0.3 = \frac{23}{8} if \ \theta = \theta_2$$

"If randomized decisions are allowed and the choice between strategies is based on expected loss only, the statistician should never take action a_3.''

2.4 Optimal Decision Rules

The fact that a best rule usually does not exist, a general method, which has been proposed for arriving at a decision rule, is frequently satisfactory.

Method of Restricting the Available Rules:

2.5 Unbiasedness

Suppose the problem is such that for each θ there exist a unique correct decision and that each decision is correct for some θ. Assume further that $L(\theta_1, d) = L(\theta_2, d)$ for all d wherever some decision is correct for both$\theta_1 and\ \theta_2$. Then the loss$L(\theta, d')$ depends only the actual decision taken, say d' and the correct decision d. thus the loss can be denoted by L(d, d') and this function measures how for a past d and d' are. Under these assumptions a decision function δ(x) is said to be unbiased w.r.t. the loss L if for all θ and d'

$$E_\theta L\big(d',\delta(x)\big) \geq E_\theta L\big(d,\delta(x)\big) \ldots\ldots\ldots\ldots\ldots \quad (3.5)$$

Where the subscript θ contains the distribution w.r.t. which the expectation can take and where d is the correct decision for θ. Thus, δ is unbiased if on the average δ(x) closer to the correct decision than to any wrong one. Extending this definition, δ is said to be L-unbiased for an arbitrary decision problem for all θ and θ'.

$$E_\theta L\big(\theta',\delta(x)\big) \geq E_\theta L\big(\theta,\delta(x)\big) \ldots\ldots\ldots\ldots \quad (3.6)$$

Example3.2: In two decision problem, let ω_0 and $\omega_1$be the set of θ values for which$d_0$andd_1 are correct decisions. Assume that

$$L(\theta, d_0) = 0 \qquad \theta \epsilon \omega_0 L(\theta, d_1) = b\theta\epsilon\omega_0 \qquad\qquad = a\theta\epsilon\omega_1$$
$$= 0 \qquad \theta\epsilon\omega_1$$

$$E_\theta L\big(\theta',\delta(x)\big) = L(\theta', d_0)P_\theta[\delta(x) = d_0] + L(\theta', d_1)P_\theta[\delta(x) = d_1]$$

$$= aP_\theta[\delta(x) = d_0] if \theta'\epsilon\omega_1$$

$$= bP_\theta[\delta(x) = d_1] if \theta'\epsilon\omega_0$$

So that (3.6) reduced to

$$aP_\theta[\delta(x) = d_0] \geq bP_\theta[\delta(x) = d_1] for \theta \epsilon \omega_0$$

With reverse inequality holding for$\theta\epsilon\omega_1$

Since$P_\theta[\delta(x) = d_0] + P_\theta[\delta(x) = d_1] = 1$ the unbiasedness contains (3.6) reduces to, $P_\theta[\delta(x) = d_1] \leq \frac{a}{a+b} for \theta\epsilon\omega_0$

And $P_\theta[\delta(x) = d_1] \geq \frac{a}{a+b} for \theta\epsilon\omega_1$

Example3.3: In the problem of estimating the real valued function $g(\theta)$ with square of the error as loss, the condition of unbiasedness become,

$E_\theta[\delta(x) - g(\theta')]^2 \geq E_\theta[\delta(x) - g(\theta)]^2$ For all θ and θ'………… (3.7)

$E_\theta[\delta(x) + E_{\theta^*}\delta(x) - E_{\theta^*}\delta(x) - g(\theta')]^2 \geq E_\theta[\delta(x) + E_\theta\delta(x) - E_\theta\delta(x) - g(\theta')]^2$

Let $E_\theta\delta(x) = h(\theta)$

$E_\theta[\delta(x) - h(\theta) + h(\theta) - g(\theta')]^2 \geq E_\theta[\delta(x) - h(\theta) + h(\theta) - g(\theta)]^2$

$[h(\theta) - g(\theta')]^2 \geq [h(\theta) - g(\theta)]^2$ For all θ and θ'

If $g(\theta)$ is continuous over Ω and which is not continuous in any open subset of Ω, and that $h(\theta) = E_\theta\delta(x)$ is continuous function of θ for each estimate$\delta(x)$ of$g(\theta)$. Thus (3.2) reduces to,

$$g^2(\theta') - 2h(\theta)g(\theta) \geq g^2(\theta) - 2h(\theta)g(\theta)$$

Or $g^2(\theta') - g^2(\theta) \geq 2h(\theta)\big(g(\theta') - g(\theta)\big)$

$$[g(\theta) - g(\theta')][g(\theta') + g(\theta)] \geq 2h(\theta)[g(\theta') - g(\theta)]$$

If θ is neither a relative minimum or maximum of $g(\theta)$ it follows that there exist points θ' arbitrary chosen θ both such that,

$g(\theta') + g(\theta) \leq 2h(\theta)$ Hence $g(\theta) = h(\theta)$

Thus $\delta(x)$ is unbiased if$E_\theta\delta(x) = g(\theta)$. <u>Proved</u>

2.6 Invariance Ordering

Generally, an invariant is a quantity that remains constant during the execution of a given operation or transformation. In other words, none of the allowed operations changes the value of the invariant. For example, any two scalar quantities the result is invariant with respect to product i.e. a×b equal b×a. In statistics this property is helpful in attempting the given problem using a more preferred form out of many available order invariant forms.

2.7 Self-Assessment Exercise

1. Discuss the decision theoretic problem as a game problem using an example from your surroundings.

2. Explain the concept of optimal Bayes rules with example.

2.8 Summery

In this unit, section 2.3 consists of the basics of Decision Theory Problem as a Game Problem and sections 2.4, 2.5 and 2.6 discuss about some Basic Elements of decision theory namely optimal decision rules, unbiasedness, and invariance ordering. In next unit we will learn more about the structures of Bayes problems.

2.9 Further Readings

4. Berger, J.O. (1985). Statistical decision theory-Fundamental concepts and methods, Springer Verlag.
5. Degroot, M. H. (1971). HPD statistical decisions, McGraw-Hill.
6. Ferguson, T.S. (1967). Mathematical statistics- A decision theoretic approach, Academic press.
7. Lindley, D.V. (1965). Introduction to probability and statistical inference from Bayesian view point, Cambridge university press.

UNIT-3: BAYES AND MINIMAX RULE

Structure

3.1 Introduction

Bayes principle refers to the notion of a distribution on the parameter space Θ called a prior distribution.

3.2 Objectives

After studying this unit, you should be able to

- Define Bayes Principle
- Define Decision rules
- Identify Minimax rules for decision theoretic problems.

3.3 Bayes and Minimax Principles

1. Bayes principle: The Bayes principle involves the notion of a distribution on the parameter space Θ called a prior distribution. Two things are needed of a prior distribution τ on Θ. First we may able to speak of the Bayes risk of a decision rule δ w.r.t. a prior distribution τ, namely

$R(\tau, \delta) = E\, R(T, \delta)$ ………………….. (3.8)

Where T is a r.v. over Θ having distribution τ. Second, we need to be able to speak of the joint distribution 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. We denote the space of prior distribution as Θ^*.

<u>Defn.</u>3.2: A decision rule δ_0 is said to be Bayes w.r.t. the prior distribution $\tau \epsilon \Theta^*$ if $R(\tau, \delta_0) = \inf_{\delta \epsilon D^*} R(\tau, \delta)$………… (3.9)

The value on the R.H.S. is known as the minimum Bayes risk. Bayes risk may not exist even if the minimum Bayes risk is defined and finite.

<u>Defn.</u>3.3: Let $\epsilon > 0$. A decision rule δ_0 is said to be $\epsilon -$ Bayes w.r.t. the prior distribution $\tau \epsilon \Theta^*$ if

$$R(\tau, \delta_0) \leq \inf_{\delta \epsilon D^*} R(\tau, \delta) + \epsilon \quad \text{……………….. (3.10)}$$

2. <u>Minimax principle</u>: An essentially different type of ordering of the decision rule 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

$$\sup_{\theta} R(\theta, \delta_1) < \sup_{\theta} R(\theta, \delta_2)$$

A rule that is most preferred in this ordering is called a minimax decision rule.

<u>Defn.</u>3.4: A decision rule δ_0 is said to be minimax if

$$\sup_{\theta \epsilon \Theta} R(\theta, \delta_0) = \inf_{\delta \epsilon D^*} \sup_{\theta} R(\theta, \delta) \quad \text{…………………… (3.11)}$$

The value on the R.H.S. of (3.11) is called the minimax value or upper value of the game.

<u>Proposition</u>3.1: A decision rule δ_0 is said to be minimax if and only if

$$R(\theta', \delta_0) \leq \sup_{\theta \epsilon \Theta} R(\theta, \delta) \quad \text{……………….. (3.12)}$$

For all θ'εΘ and δεD^*

<u>Proof</u>: let $R(\theta', \delta_0) \leq \sup_{\theta \epsilon \Theta} R(\theta, \delta)$ For all θ'εΘ and δεD^*

$\sup_{\theta' \epsilon \Theta} R(\theta', \delta_0) \leq \sup_{\theta \epsilon \Theta} R(\theta, \delta)$ for δεD^*

Hence δ_0 minimizes the $\sup_{\theta \epsilon \Theta} R(\theta, \delta)$ for δεD^*

Thus, $\sup_{\theta'\epsilon\Theta} R(\theta',\delta_0) = \inf_{\delta\epsilon D^*}\sup_{\theta\epsilon\Theta} R(\theta,\delta)$ And δ_0 is minimax.

Conversely, let $\sup_{\theta\epsilon\Theta} R(\theta,\delta_0) = \inf_{\delta\epsilon D^*}\sup_{\theta\epsilon\Theta} R(\theta,\delta)$

$$\Rightarrow \sup_{\theta\epsilon\Theta} R(\theta,\delta_0) \le \sup_{\theta\epsilon\Theta} R(\theta,\delta) \text{ for } \delta\varepsilon D^*$$

$$\Rightarrow R(\theta',\delta_0) \le \sup_{\theta\epsilon\Theta} R(\theta,\delta_0) \le \sup_{\theta\epsilon\Theta} R(\theta,\delta) \text{ for all } \theta'\varepsilon\Theta,\ \delta\varepsilon D^*$$ Proved

Defn.3.5: Let $\epsilon > 0$. A decision rule δ_0 is said to be ϵ - minimax if

$$\sup_{\theta\epsilon\Theta} R(\theta,\delta_0) \le \inf_{\delta}\sup_{\theta\epsilon\Theta} R(\theta,\delta) + \epsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.13)$$

More simply, δ_0 is ϵ-minimax if for all θ'εΘ and $\delta\varepsilon D^*$

$$R(\theta',\delta_0) \le \sup_{\theta\epsilon\Theta} R(\theta,\delta) + \epsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.14)$$

Defn.3.6: A distribution $\tau_0\,\varepsilon\,\theta^*$ is said to be *least favorable* if $\inf_{\delta} \gamma(\tau_0,\delta) = \sup_{\tau}\inf_{\delta} \gamma(\tau,\delta)$ ……………….. (3.15)

The value on the R.H.S. of (3.15) is called the maximin value or lower value of the game.

Geometrical Interpretation for finite Θ: we give a geometric interpretation of the fundamental problem of decision theory in the case in which the parameter space Θ is finite.

Suppose that Θ contains k points, $\Theta = \{\theta_1,\theta_2,\ldots\ldots,\theta_k\}$ and consider the set S, to be called the *risk set*, contained in k-dimensional Euclidian space E_k of points of the form $\big(R(\theta_1,\delta),R(\theta_2,\delta),\ldots\ldots\ldots\ldots,R(\theta_k,\delta)\big)$, where δ ranges through D^*

$$S = \left\{(y_1,y_2,\ldots\ldots,y_k)\, for\ some\ \delta\epsilon D^*, y_j = R\big(\theta_j,\delta\big)\ for\ j = 1,2,\ldots,k\right\}$$

………… (3.16)

If k=2 this set may easily be plotted in the plane.

Defn.3.7: A set S should be convex if when ever $y = (y_1,y_2,\ldots\ldots,y_k) y' = (y'_1,y'_2,\ldots\ldots,y'_k)$ are elements of S, the point

$\alpha y + \overline{1-\alpha}y' = (\alpha y_1 + \overline{1-\alpha}y'_1,\ldots\ldots,\alpha y_k + \overline{1-\alpha}y'_k)$ are also elements of S , $0 \le \alpha \le 1$.

Lemma3.1: The risk set S is convex subset of E_k.

<u>Proof</u>: Let y and y' be arbitrary point of S. according to the definition of S, there exist a decision rules δ and δ' in D^* for which $y_j = R(\theta_j, \delta)$

and $y_j' = R(\theta_j, \delta')$ j=1, 2,....., k .let α be an arbitrary number such that $0 \le \alpha \le 1$ and consider $\delta_\alpha = \alpha\delta + \overline{1-\alpha}\delta'$. Clearly $\delta_\alpha \epsilon\, D^*$. (as convex combination of d.f is also a d.f)

$$R(\theta_j, \delta_\alpha) = E\ L(\theta_j, \delta_\alpha) = \alpha\ E\ L(\theta_j, \delta) + \overline{1-\alpha} E\ L(\theta_j, \delta')$$

$$= \alpha R(\theta_j, \delta) + \overline{1-\alpha} R(\theta_j, \delta') = Z_j$$

$Z = (Z_1, Z_2, \dots \dots, Z_k)\ \epsilon\ S$ <u>Proved</u>

<u>Defn.</u>3.8: let A be a set. The convex hull of a set A is the smallest convex set containing A or the intersection of all convex sets containing A.

Thus S defined above is the convex hull of the set S_0, where

$$S_0 = \{(y_1, y_2, \dots \dots, y_k)\ \ y_j = R(\theta_j, d), d\epsilon D, j = 1,2, \dots, k\} \text{ .... (3.17)}$$

Because the risk function contains all the pertinent information about a decision rule as for as we concerned, the risk set S contains all the information about a 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-dimensional space there is a decision problem, (Θ, D^*, R) in which Θ consists of k points, whose risk set is the set S.

<u>Bayes Rules</u>:

let $(p_1, p_2, \dots \dots, p_k)$ be a probability distribution on Θ. See points that yield the same expected risk.

$\sum_{j=1}^{k} p_j\ R(\theta_j, \delta) = \sum p_j\ y_j$,$y_j = (\theta_j, \delta)$ .... (3.18)

are equivalent in the ordering given by the principle for the prior distribution $(p_1, p_2, \dots \dots, p_k)$. Thus all points on the plane $\sum p_j\ y_j = b$ for any real number b are equivalent. Every such plane is perpendicular to the vector from the origin to the points $(p_1, p_2, \dots \dots, p_k)$ and because p_j is non negative the slope of the line of the interaction of the plane $\sum p_j\ y_j = b$ with the coordinate planes cannot be positive. The quantity b can best be visualized by noting that the point of interaction of the diagonal line $y_1 = y_2 = .. = y_k$ with the plane $\sum p_j\ y_j = b$ must occur at $(b, b, \dots, b)$

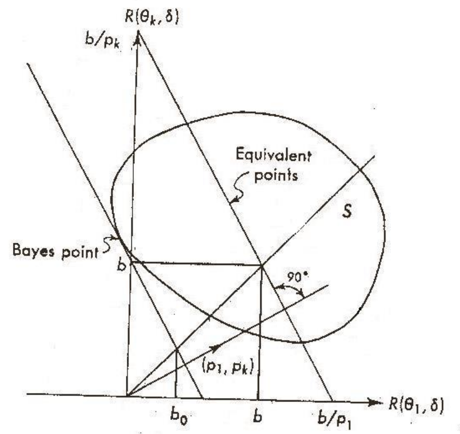


Fig (3.1)

To find the Bayes rules we find the infimum of those values of b, call it b_0, for which the plane $\sum p_j\, y_j = b$ intersected the set S. decision rule corresponding to points in the intersection are Bayes rule with respect to the prior distribution $(p_1, p_2, \dots \dots, p_k)$. There may be many Bayes rules or there may not be any Bayes rules.

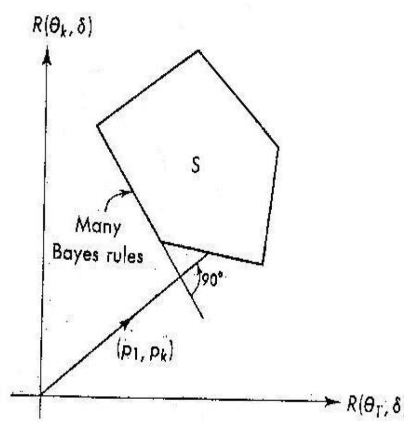


Fig (3.2)

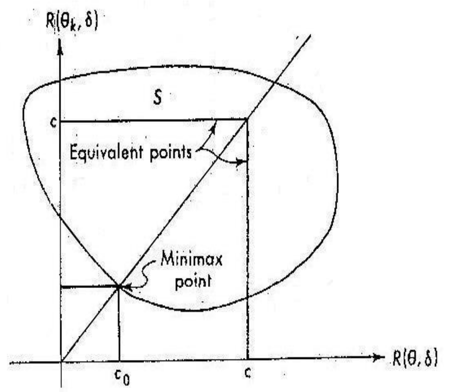


Fig (3.3)

Minimax Rules:

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

$Q_c = \{(y_1, y_2, \dots \dots, y_k): y_j \leq c \quad for\ i = 1, \dots \dots, k\}$for any real number 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 intersects S. any decision rule δ, whose associated risk point is an element of the intersection $Q_{c0} \cap S$, is minimax decision rule. Of course, minimax decision rule do not exist when the set S does not contains its boundary points.

A minimax strategy for nature which is otherwise called a ''least favorable distribution'' may also be visualized geometrically. A strategy for nature is a prior distribution $\tau = (p_1, p_2, \dots \dots, p_k)$ Because the minimum Bayes risk $inf_\delta \Upsilon(\tau, \delta)$ isb_0, where $(b_0, b_0, \dots \dots, b_0)$ in the intersectionof the line $y_1 = y_2 = \cdots = y_k$ and the plane, tangent to and below S, and perpendicular to $(p_1, p_2, \dots \dots, p_k)$, a least favorable distribution is the choice of $(p_1, p_2, \dots \dots, p_k)$ that makes this intersection as for up the line aspossible. It is clear thatb_0 is not greater thanc_0, the minimax risk is c_0.This distribution must be least favorable.

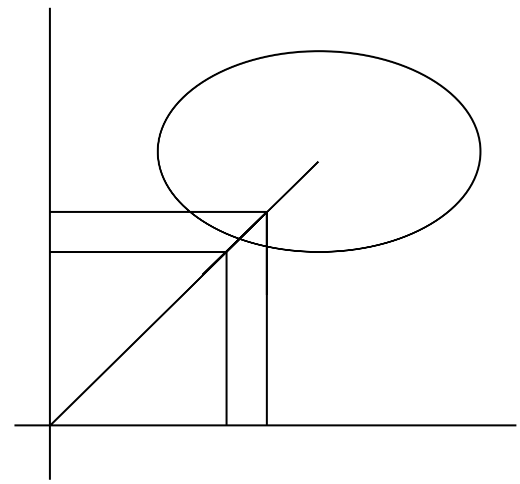


Fig (3.4)

Since

$R(\theta,\delta) = E\ R(\theta,Z)$where Z is a r.v. taking values in D with d.f δ.

$if\ \delta_0$is such that $R(\theta,\delta_0) = inf_{\delta\varepsilon\,\mathrm{D}^*} R(\theta,\delta)$ then

$R(\theta,\delta_0) = E\ R(\theta,Z)$where Z is a r.v. taking values in D with d.f δ_0.

Obviously $\int R(\theta,\delta_0)d\tau \leq \int R(\theta,d)d\tau$ for all dε D

$\Upsilon(\tau,\delta_0) = \int R(\theta,\delta_0)d\tau \leq inf_{\mathrm{d}\varepsilon\,\mathrm{D}}\Upsilon(\tau,d)$

$\Upsilon(\tau,\delta_0) = inf_{\delta\varepsilon\,\mathrm{D}^*}\Upsilon(\tau,\delta) \leq inf_{\mathrm{d}\varepsilon\,\mathrm{D}}\Upsilon(\tau,d)$ ………………… (3.19)

Also $R(\theta,\delta_0) = E\ R(\theta,Z)$ Z is a r.v. taking values in D with d.f δ_0.

$$= \int R(\theta,Z)d\delta_0$$

$$\int R(\theta,\delta_0)d\tau = \int \left[\int R(\theta,Z)d\delta_0\right] d\tau$$

$$= \int \left[\int R(\theta,Z)d\,\tau\right] d\,\delta_0$$

$\Upsilon(\tau,\delta_0) = \int[\int R(\theta,Z)d\,\tau]d\,\delta_0$

$$\geq \int \left[inf_{\mathrm{d}\varepsilon\,\mathrm{D}} \int R(\theta,Z)d\,\tau\right] d\,\delta_0$$

$= inf_{\mathrm{d}\varepsilon\,\mathrm{D}}\Upsilon(\tau,d)$

$\Upsilon(\tau,\delta_0) \geq inf_{\mathrm{d}\varepsilon\,\mathrm{D}}\Upsilon(\tau,d)$ ………………….. (3.20)

From (4.19) and (4.20)

$$\Upsilon(\tau,\delta_0)=inf_{\mathrm{d\varepsilon\ D}}\Upsilon(\tau,d) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.21)$$

Equation (3.21) states that none of the mixed strategy (randomized decision rule) can reduce the risk below the minimum value which can be attained from the non-randomized decision D. if Bayes risk $\Upsilon(\tau,\delta_0)$ is finite and is attained for a randomized decision rules δ_0, then it follows from the above comments that this risk must be attained for some non- randomized decision D.

Thus if a Bayes rule with respect to a prior distribution τ exits,there exist a non- randomized Bayes rule w.r.t. τ. Therefore, one definite computational advantage that the Bayes approach has over the minimax approach to decision theory problem is that the search for good decision rules may be restricted to the class of non- randomized decision rules.

<u>Example</u>3.4: Let $\Theta = a = \{0,1\}$ and let the loss function be L(0,0) = L(1,1) = 0 , L(1,0)=L(0,1)=1 Suppose that the statistician observes the r.v. X with discrete distribution

$$P[X=x/\theta]=2^{-K} \quad K=x+\theta \quad k=1,2,3,\ldots\ldots\ldots$$

(I) Describe the set of all non- randomized decision rules.

(II) Plot the risk set S in the plane.

(III) Find the minimax and Bayes decision rules.

<u>Sol</u>: $\mathfrak{X}=N=$ set of all non- negative integers

Let A be any finite subset of N. d:$\mathfrak{X}\rightarrow a$ = {0,1}

$$D=\{d: \quad d:\mathfrak{X}\rightarrow a\}$$

Thus D contains only two types of functions

$$d_1(x)=1 \quad if\ x\varepsilon A \qquad d_2(x)=1 \quad if\ x\varepsilon A'$$

$$=0 \quad if\ x\varepsilon A' \qquad =0 \quad if\ x\varepsilon A$$

The cardinality of D is C

$R(\theta,d)=E\ L\big(\theta,d(X)\big)$is risk function of d.

$$R(0,d_1)=E\ L\big(0,d_1(X)\big)=P[X\varepsilon A] \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.22)$$

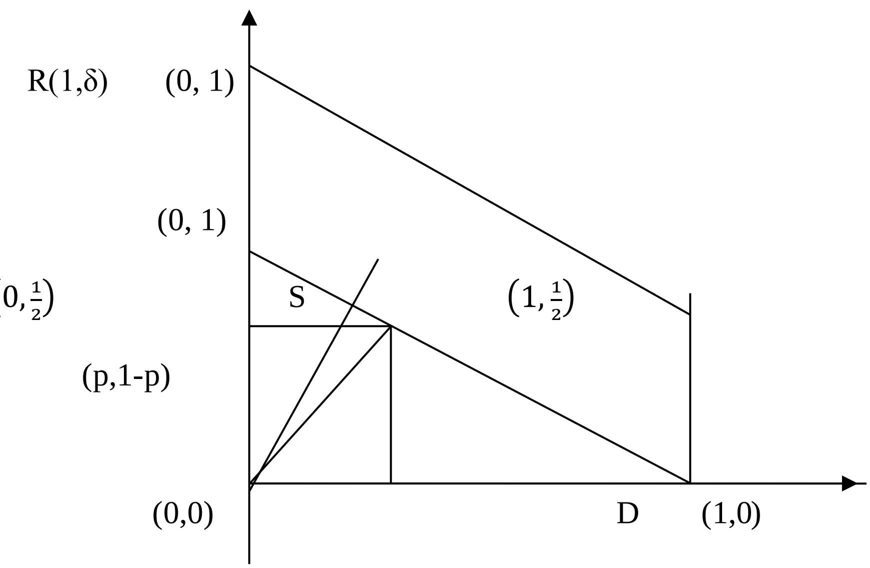
$$R(1,d_1)=E\ L\big(1,d_1(X)\big)=P[X\varepsilon A'] \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.23)$$

$$R(0,d_2)=E\ L\big(0,d_2(X)\big)=P[X\varepsilon A'] \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.24)$$

$R(1, d_1) = E\, L(1, d_2(x)) = P[X \varepsilon A]$ ……………………. (3.25)

$R(\theta, \delta) = \int R(\theta, Z) d\delta$ Where Z is a r.v. taking values in D with d.f δ.

Let A ={0} ,{0,1}, Φ



$R(0, d_1) = P[X \varepsilon A] = 0, 1/2, 0$

$R(0, d_2) = P[X \varepsilon A'] = 1, 1/2, 1 \quad y_2$

$R(1, d_1) = P[X \varepsilon A'] = 1/2, 1/4, 1$

$R(1, d_2) = P[X \varepsilon A] = \frac{1}{2}, \frac{3}{4}, 0 \qquad L_2\left(0, \frac{1}{2}\right)$

(0,1/2) , (1/2,1/4) , (0,1)

(1,1/2) , (1/2,3/4) , (1,0)

$S = \{(\alpha, \beta): 0 \leq \alpha \leq 1, 0 \leq \beta \leq 1\}$

$L_1 y_1 =$ R(0,δ), $y_2 = R(1, d)$

$\alpha = R(0, d), \beta = R(1, d) \qquad d \varepsilon D$Fig (3.5)

Thus minimax decision rule δ_0 at point D

i.e line $L_1\, L_2$ and intersection of $y_1 = y_2$

Line $L_1\, L_2$ is $2y_2 + y_1 = 1$

Where $y_1 = y_2 \Rightarrow D = (\frac{1}{3}, \frac{1}{3})$

So corresponding to$\left(\frac{1}{3}, \frac{1}{3}\right)$ is $\left(\frac{2}{3}, \frac{1}{3}\right)$.

A Bayes decision rule which minimize (3.23) can be found.

<u>To find a non-randomized rule</u>:

Let A= {1, 3, 5, 7 …} $\quad d(x) = \begin{matrix} 0 & x \epsilon A \\ 1 & x \epsilon A' \end{matrix}$

$$R(0, d) = EL(0, d) = P[X \epsilon A'] = \sum_{x=2,4,6,\ldots} 2^{-x}$$

$$= \frac{1}{2^2} + \frac{1}{2^4} + \cdots = \frac{\frac{1}{4}}{1 - \frac{1}{4}} = \frac{1}{3}$$

$$R(1,d) = EL(1,d) = P[X\epsilon A] = \sum_{x=1,3,5,\ldots} 2^{-(x+1)}$$

$$= \frac{1}{2^2} + \frac{1}{2^4} + \cdots = \frac{\frac{1}{4}}{1-\frac{1}{4}} = \frac{1}{3}$$

Thus there exist a non-randomized Bayes decision rule such that$\left(\frac{1}{3},\frac{1}{3}\right) = point\ D\ with\ probability\ \left(\frac{2}{3},\frac{1}{3}\right)$. A minimax decision rule is$\left(\frac{2}{3},\frac{1}{3}\right)$ choosing,

$$\mathrm{d_1(x) = 0 \text{ if } x = 0 \text{with probability} } \frac{2}{3}\text{and}$$

=1 if x$\geq$ 1

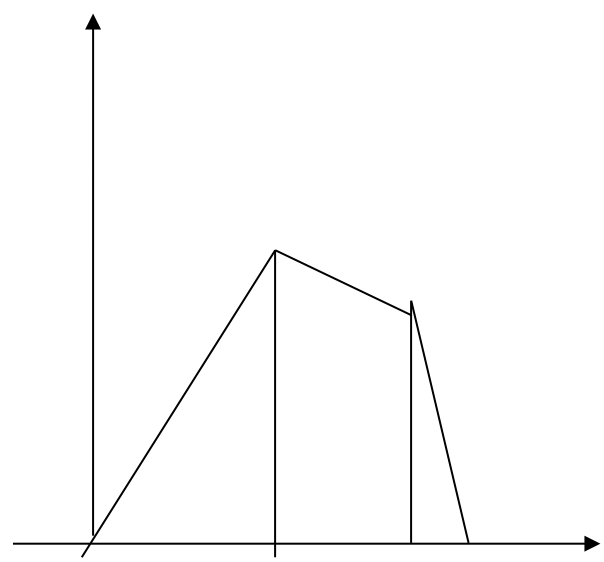
$$\mathrm{d_2(x) = 1 \quad x \geq 0 \text{ with probability} } \frac{1}{3}$$

This rule is also Bayes rule with $(p_1,p_2) = \left(\frac{1}{3},\frac{2}{3}\right) = (p,1-p)\ as$

$$\frac{1-p}{p}\left(-\frac{1}{2}\right) = -1 \Rightarrow 2p = 1-p \Rightarrow p = \frac{1}{3}$$

<u>Example</u>3.5: consider the statistical decision problem.

$$\Omega = (\theta_1,\theta_2)\ D = (d_1,d_2) L(\theta,d)\ as$$



$L(\theta,d)$		d_1	d_2	$\rho^*(\alpha)$
	θ_1	0	a_1	$a_i > 0\ \ i = 1,2$
	θ_2	a_2	0	

Let $\alpha(\delta) = P[\delta(x) = d_2/\theta = \theta_1]$

$and \quad \beta(\delta) = P[\delta(x) = d_1/\theta = \theta_2]\frac{8}{17}\frac{16}{17}\alpha$

Fig (3.8)

$\alpha(\delta)\ and\ \beta(\delta)$ are the probabilities

that δ will lead to a decision when $\theta = \theta_1$ and $\theta = \theta_2$ respectively, suppose $P[\theta = \theta_1] = \xi$

$P[\theta = \theta_2] = 1 - \xi$, 0< $\xi < 1$ is the prior probability.

$$\Upsilon(\tau, \delta) = \iint L(\theta, \delta)\, dF(x/\theta) d\tau(\theta)$$

$$= \int \{L(\theta, d_1)\, P[\delta(x) = d_1/\theta] + L(\theta, d_2) P[\delta(x) = d_2/\theta]\} d\tau(\theta)$$

$= \big[L(\theta_1, d_1) P[\delta(x) = d_1/\theta_1] + L(\theta_1, d_2) P[\delta(x) = d_2/\theta_1]\big]\xi$

$+\big[L(\theta_2, d_1) P[\delta(x) = d_1/\theta_2] + L(\theta_2, d_2) P[\delta(x) = d_2/\theta_2]\big](1 - \xi)$

$= L(\theta_1, d_2) P[\delta(x) = d_2/\theta_1]\xi + L(\theta_2, d_1) P[\delta(x) = d_1/\theta_2](1 - \xi)$

$= a_1 \alpha(\delta)\xi + a_2 \beta(\delta)(1 - \xi)$

$= a\alpha(\delta) + b\beta(\delta)$…….. (3.33)   Where, $a = a_1\xi$, b $= a_2(1 - \xi)$

<u>**Example**</u>**3.6:** $\Theta = \{\theta_1, \theta_2\}$ a $= \{a_1, a_2\}$

$$L(\theta, a) = \begin{array}{c|cc} & a_1 & a_2 \\ \hline \theta_1 & -2 & 3 \\ \theta_2 & 3 & -4 \end{array}$$

A randomized strategy $\delta \epsilon$a* is represented as a number $0 \le q \le 1$, with understanding that $a_1$is taken with probability q and a_2 with 1-q

$$S = \{(L(\theta_1, \delta), L(\theta_2, \delta)), \delta \epsilon \text{a}^*\}$$

$$L(\theta_1, \delta) = EL(\theta_1, z) = L(\theta_1, a_1) P_{\theta_1}[z = a_1] + L(\theta_1, a_2) P_{\theta_1}[z = a_2]$$

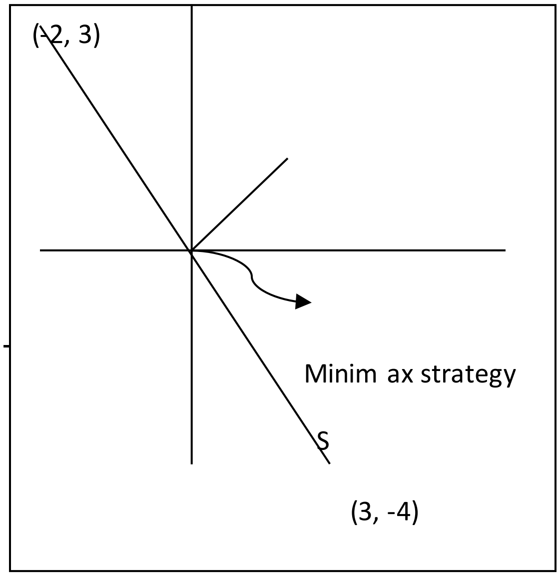
$$= -2q + 3(1 - q) = 3 - 5q$$

Similarly, $L(\theta_2, \delta) = EL(\theta_2, z) = 3q - 4(1 - q) = 7q - 4$

$S = \{(3 - 5q, 7q - 4), 0 \le q \le 1\}$ (Fig 3.6)

Which is nearly a line segment joining (-2, 3) and (3, -4) minimax strategy occurs when,

$3-5q=7q-4 \quad or\ q=\frac{7}{12}$



The minimax risk is $\left(\frac{1}{12},\frac{1}{12}\right)$

Thus, minimax rule is $\left(\frac{7}{12},\frac{5}{12}\right)$

And this is also Bayes rule since,

$\frac{1-p}{p}\left(-\frac{7}{5}\right)= -$

$If\ we\ choose\ \theta_1\ with\ prob.\frac{7}{12}$

$And\theta_2\ with\ prob.\frac{5}{12}.\left(\frac{7}{12},\frac{5}{12}\right)\ is\ prior\ probability.$

<u>Example</u>3.7: $\Theta=\{1,2\}=$ a

$$\begin{aligned}
d_1(1)&=1 \quad ,d_1(1)=1\\
d_2(1)&=1 \quad ,d_2(1)=2\\
d_3(1)&=2 \quad ,d_3(1)=1\\
d_4(1)&=2 \quad ,d_4(1)=2
\end{aligned}$$

	d_1	d_2	d_3	d_4
1	-2	$-\frac{3}{4}$	$\frac{7}{4}$	3
2	3	$-\frac{9}{4}$	$\frac{5}{4}$	-4

$$R(\theta_1,\delta)=p_1R(\theta_1,d_1)+p_2R(\theta_1,d_2)+p_3R(\theta_1,d_3)+p_4R(\theta_1,d_4)$$

$$=-2p_1-\frac{3}{4}p_2+\frac{7}{4}p_3+3p_4,\sum p_i=1$$

$$R(\theta_1,\delta)=\sum_{i=1}^{4}p_iR(\theta_2,d_i)=3p_1-\frac{9}{4}p_2+\frac{5}{4}p_3-4p_4$$

$S=\left\{\left(R(\theta_1,\delta),R(\theta_2,\delta)\right):\delta\epsilon\text{a}^*\right\}$(Fig 3.7)

Line L_1L_2 is $y_2 = -\frac{21}{5}y_1 - \frac{27}{5}$

$$5y_2 + 21y_1 + 27 = 0$$

Line PQ intersects L_1L_2 at

$y_1 = -\frac{27}{26}, y_2 = (-27)/26$ Thus

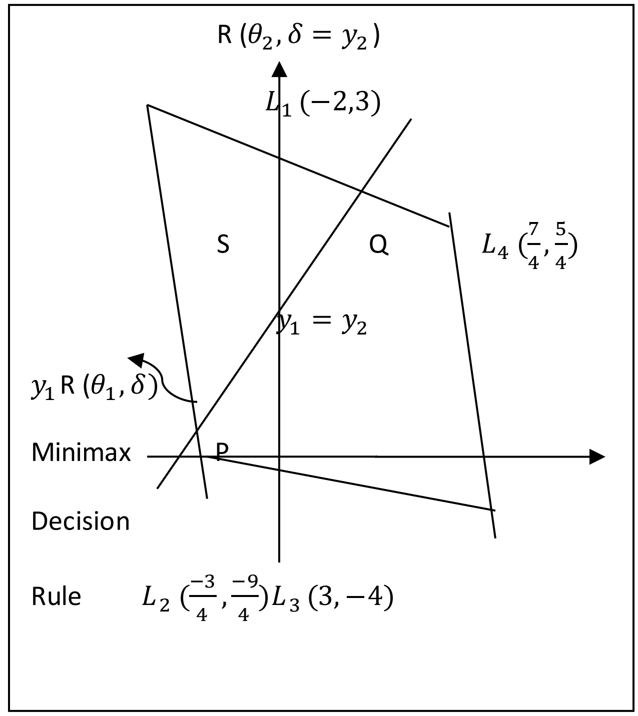
The Minimax risk at $\left(\frac{-27}{26}, \frac{-27}{26}\right)$

Thus δ_0 corresponding to this

Minimum is attained by

$$\delta_0 = \left(\frac{3}{13}, \frac{10}{13}, 0,0\right).$$

Thus δ_0 is also bayes w.r.to



$$\left(\frac{21}{26}, \frac{5}{26}\right) = \tau \quad as\, \frac{1-p}{p}\left(-\frac{21}{5}\right) = -1 \Rightarrow (1-p)21 = 5p \Rightarrow p = \frac{21}{26}$$

And minimum Bayes risk $\gamma(\tau, \delta_0) = \frac{21}{26}$

Also d_1 is non- randomized bayes rule w.r.to τ as

$$\gamma(\tau, d_1) = pR(\theta_1, d_1) + (1-p)R(\theta_2, d_1)$$

$$= \frac{21}{26}(-2) + \frac{5}{26}(3) = \frac{-42+15}{26} = -\frac{27}{26}$$

Thus $\delta_0 = \left(\frac{3}{13}, \frac{10}{13}, 0,0\right)$is randomized Bayes rule and d_1 is non-randomized Bayes rule w.r.to $\tau = \left(\frac{21}{26}, \frac{5}{26}\right)$

Thus, minimax Bayes risk is $-\frac{27}{26}$.

Given the prior distribution τ, we want to choose a non –randomized decision rule dϵD that minimizes Bayes risk,

$$\gamma(\tau, d) = \int R(Z, d)d\tau \qquad where,\, Z\, is\, a\, random\, variable\, taking\, values$$

$$R(\theta, d) = \int L(\theta, d(x))\, dF_X(x/\theta)$$

A choice of θ by the distribution τ (θ), 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,

$$F_X(x) = \int F_X(x/\theta) d\tau(\theta) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.26)$$

and then choosing θ according to the conditional distribution of θ, given X=x, $\tau(\theta/x)$. Hence by a change of integration we may write,

$$\gamma(\tau, d) = \int [\int L(\theta, d(x))\, \mathrm{d}\tau(\theta/x)]\, dF_X(x) \quad \ldots\ldots\ldots\ldots\ldots \quad (3.27)$$

Given that these operations are legal, it is easy to describe aBayes decision rule.

To find a function d(x) that minimizes the double integral (3.27), we may minimize the inside integral separately for each x; that is, we may find for each x the action, call it d(x), that minimizes

$$\int L(\theta, d(x))\, \mathrm{d}\tau(\theta/x)$$

Thus, the Bayes decision rule minimizes the posterior conditional expected loss, given the observations.

Non–Negative Loss Function:

Suppose that the distribution of the parameter θ in some decision problem is τ (θ). Let a be a given constant (>0), and let $\lambda(\theta)$ be a real valued function over parameter space Θ=Ω, such that

$$\int_{\Omega} \lambda(\theta) d\tau(\theta) < \infty$$

Consider a new loss function L_0 which is defined in terms of the original loss function L by relation

$$L_0(\theta, d) = aL(\theta, d) + \lambda(\theta) \qquad \theta\epsilon\Omega\ , d\epsilon D \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.28)$$

For any decision dϵD, let Y (τ, d) denote the risk which results from the original loss function L.

$$\gamma(\tau, d) = \int R(\theta, d) d\tau = \int \int L(\theta, d)\, dF(x/\theta)\, d\tau(\theta) \quad \ldots\ldots\ldots \quad (3.29)$$

And let $\gamma_0(\tau, d) = \int \int L_0(\theta, d)\, dF(x/\theta)\, d\tau(\theta)$ …………… (3.30)

Then for any two decisions $d_1 and\ d_2 \epsilon D$

$\gamma_0(\tau, d_1) \leq \ \gamma_0(\tau, d_2) \Leftrightarrow \Upsilon(\tau, d_1) \leq \Upsilon(\tau, d_2)$ …………….. (3.31)

In particular, a decision d* is Bayes w.r.to τ in the original problem with loss function L (θ, d) if and only if d* is a Bayes w.r.to τ in the new problem with loss function L_0.

Now consider $\lambda_0(\theta) = \inf_{d\epsilon D} L(\theta, d)$

$If \int_\Omega \lambda_0(\theta) d\tau(\theta) < \infty$, We can replace L now by a new loss function L_0 which is defined as,

$$L_0(\theta, d) = L(\theta, d) - \lambda_0(\theta)$$

Then loss function L_0 has the following property

$$\left.\begin{array}{l} L_0(\theta, d) \geq 0 \quad for\ all\ \theta\ and\ d\ and \\ \inf_{d\epsilon D} L_0(\theta, d) = 0 \end{array}\right\} \quad \ldots\ldots\ldots\ldots.. \ (3.32)$$

It has been found convenient in many problems to role with non-negative loss function of this type, although the use of such function makes it appear that the statistician must continually choose decisions from which he can never realize a positive gain.

3.4 Generalized Bayes Rules and Extended Bayes Rules

<u>Defn.</u>3.9: A rule δ is said to be limit of Bayes rules δ_n, if for almost all x

$\delta_n(x) \longrightarrow \delta(x)$ (In the sense of distribution) for non-randomized decision rules this definition becomes $d_n \longrightarrow d$ if $d_n(x) \longrightarrow d(x)$ for almost all x.

<u>Def 3.10</u>: A rule δ_0 is said to be generalized Bayes rules if there exist a measure τ on Θ (or non decreasing function on θ if Θ is real), such that $R(\tau, \delta) = \int\int L(\theta, \delta) f(x/\theta)\, d\tau(\theta)$ takes on a finite minimum value when $\delta = \delta_0$

<u>Def 3.11</u>: A rule δ_0 is said to be extended Bayes rules if δ_0 is ϵ- Bayes for every $\epsilon > 0$.

In other words, δ_0 is extended Bayes rules if for every $\epsilon > 0$ there exist a prior distribution τ such that δ_0 is ϵ- Bayes w.r.to τ i.e

$$\Upsilon(\tau, \delta_0) \leq \inf_{\delta} \Upsilon(\tau, \delta)$$

<u>Example</u>3.8: let $X \sim N\ (\theta, 1)$ and let $\tau(\theta) = N(0, \sigma^2)$

$L(\theta, d) = (\theta - d)^2$ The joint p.d.f of (θ, x)

$$h(\theta, x) = \frac{1}{2\pi\sigma} \exp\left[\frac{-(x-\theta)^2}{2} - \frac{\theta^2}{2\sigma^2}\right]$$

$$f_X(\mathrm{x}) = \frac{1}{2\pi\sigma} \int \exp\left[\frac{-(x-\theta)^2}{2} - \frac{\theta^2}{2\sigma^2}\right] d\theta$$

$$= [2\pi(1+\sigma^2)]^{\frac{-1}{2}} \exp\left[\frac{x^2}{2(1+\sigma^2)}\right]$$

Posterior density of θ given x,

$$f(\theta/x) = \frac{(1+\sigma^2)^{\frac{-1}{2}}}{(2\pi\sigma^2)^{\frac{-1}{2}}} \exp\left[\frac{-1+\sigma^2}{2\sigma^2}\left(\theta - \frac{x\sigma^2}{1+\sigma^2}\right)^2\right]$$

$$\sim N\left(\frac{x\sigma^2}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right)$$

The Bayes rule w.r.to τ_σ is posterior mean i.e $d_\sigma(\mathrm{x}) = \frac{x\sigma^2}{1+\sigma^2}$

The Bayes risk, $\Upsilon(\tau_\sigma, d_\sigma) = E[E(\theta - d_\sigma(x))^2/X] = \frac{\sigma^2}{1+\sigma^2}$

Thus d(x)=x is not Bayes.

But $d_\sigma(\mathrm{x}) \longrightarrow \mathrm{d(x)}$ as $\sigma \longrightarrow \infty$.

<u>Theorem 3.1</u>: for any constants a, b >0, let δ^* be a decision rule such that $\delta^*(x) = d_1 \quad if\ af_1(x) > bf_2(x)$

$$= d_2 \quad if\ af_1(x) < bf_2(x)$$

where f_i denote the conditional p.d.f of X for $\theta = \theta_i$, $i = 1,2$

The value of $\delta^*(x)$ may be either $d_1\ or\ d_2$ if $af_1(x) = bf_2(x)$. Then for any other decision function δ we have

$$a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta)$$

<u>Proof</u>: let $S_1 = \{x: \delta(x) = d_1\}$, $S_2 = \{x: \delta(x) = d_2\} = {S_1}^c$

$$A = \{x: af_1(x) > bf_2(x)\} \qquad B = \{x: af_1(x) < bf_2(x)\}$$

Then $\quad a\alpha(\delta) + b\beta(\delta) = a\int_{S_2} f_1 d\mu + b\int_{S_1} f_2 d\mu$

$= a + \int_{S_1} (bf_2 - af_1)d\mu$ ………………. (3.34)

(3.34) will be minimum if $\int_{S_1}(bf_2 - af_1)d\mu < 0$

Thus $a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta)$.

Finding a decision function δ which minimize the linear combination

$a\alpha(\delta) + b\beta(\delta)$is equivalent to finding a set S_1 for which the integral

$\int_{S_1}(bf_2 - af_1)d\mu$is minimized. This integral will be minimized if the set S_1includes every point x ε S (sample space) for which the integral is negative and excludes every point x ε S for which the integral is positive.

Remark: the posterior distribution of $\theta = \theta_1$ given X=x , denoted as α(x) is given by,

$$\alpha(x) = P[\theta = \theta_1/X = x]$$

$$= \lim_{h\to 0}\frac{P[\theta = \theta_1, x - h < X \leq x + h]}{P[x - h < X \leq x + h]}$$

$$= \lim_{h\to 0}\frac{P[x - h < X \leq x + h/\theta = \theta_1]P(\theta = \theta_1)}{P[x - h < X \leq x + h]}$$

$$= \frac{f(x/\theta_1)P(\theta = \theta_1)}{f_x(x)} = \frac{\alpha f(x/\theta_1)}{f_x(x)} = \frac{\alpha f_1(x)}{\alpha f_1(x) + 1 - \alpha f_2(x)}$$

Provided limit exists, where

$$f_1(x) = f(x/\theta_1)\ , f_2(x) = f(x/\theta_2)$$

Posterior risk of $d_1 = L(\theta_1, d_1)\alpha(x) + L(\theta_2, d_1)\big(1 - \alpha(x)\big)$

$$= a_2\big(1 - \alpha(x)\big) \qquad \text{Similarly, } d_2 = a_1\alpha(x)$$

We choose d_2 if (i.e d_2 is Bayes rule) posterior risk of $d_2 <$ posterior risk of d_1. i.e

$$a_1\alpha(x) < a_2\big(1 - \alpha(x)\big)\ or\ a_1\alpha f_1(x) < a_2\overline{1 - \alpha}f_2(x)$$

Thus $\delta^*(x) = d_2(x)\ if\, a_1\alpha f_1(x) < a_2\overline{1 - \alpha}f_2(x)$

Let $S_2 = \left\{x: \frac{f_2(x)}{f_1(x)} > \frac{a_1\alpha}{a_2(1-\alpha)}\right\}$ then, $\delta^*(x) = d_2(x) \qquad if\, x\varepsilon S_2$

$$= d_1(x) \qquad if\, x\varepsilon {S_2}^c$$

For testing $H_0: \theta = \theta_1 against\ H_1: \theta = \theta_2$,

$$d_1 = acceptH_0\ , d_2 = rejectH_0\ ,$$

$\delta^*(x) = \{0,1\}$i.e choosing d_1 with prob. 0 and d_2 with prob.1.

Or $\delta^*(x) = 1 \quad if\ x\varepsilon S_2$

$$= 0 \quad if\ x\varepsilon {S_2}^c$$

For each θ we have a d.f. of r.v. X as F(x/θ). Let G (θ) is the d.f. of r.v. θ. Then,

$$F(x/\theta) = \lim_{k\to 0} \frac{P[X\le x, \theta-K<\theta<\theta+K]}{P[\theta-K<\theta<\theta+K]} = \lim_{k\to 0} \frac{\int_{-\infty}^{x}\int_{\theta-k}^{\theta+k} f(t,v)dtdv}{\int_{\theta-k}^{\theta+k} f_\theta(v)dv}$$

Provided such $f(t,v)$,$f_\theta(v)$ exist and also limit exists. If $f(t,v)$ and $f_\theta(v)$ are continuous.

$$F(x/\theta) = \lim_{k\to 0} \frac{2K\int_{-\infty}^{x} f(t,v_0)dt}{2Kf_\theta(v_0)} \qquad \text{Where } v_0\varepsilon(\theta - k, \theta + k)$$

$$= \frac{\int_{-\infty}^{x} f(t,\theta)dt}{f_\theta(\theta)}$$

Since $f(t,v)$ is assumed to be continuous, then

$$F(x/\theta) = \frac{f(x,\theta)}{f_\theta(\theta)} = \frac{f(x,\theta)}{g(\theta)} \qquad g(\theta) = f_\theta(\theta)$$

Similarly, $F(x/\theta) = \lim_{k\to 0} \frac{P[X\le x, \theta-K<\theta<\theta+K]}{P[\theta-K<\theta<\theta+K]} = \ = \frac{\int_{-\infty}^{x} f(x,v)dv}{f_X(x)}$

The posterior density of θ given x (when observation X=x is taken.)

$$F(x/\theta) = \frac{f(x,\theta)}{f_\theta(\theta)} = \frac{f(x,\theta)}{\int f(x,\theta)d\theta} = \frac{F(x/\theta)g(\theta)}{\int f(x/\theta)g(\theta)d\theta}$$

This is a continuous version of Bayes theorem.

3.5 Limit of Bayes Rules

the limiting Bayes method): Suppose $\bar{X}$ is not admissible, and without loss of generality we may assume σ=1. Then there exists δ^* such that

$$\left.\begin{array}{l} R(\theta,\delta^*) \le \frac{I}{n} \;\; for\ all\ \theta \\ \quad < \frac{I}{n}\ for\ some\ \theta \end{array}\right\} \ (under\ the\ square\ error\ loss\ function)$$

R (θ, δ) is a continuous function of θ for every δ, so that there exist

$\varepsilon > 0\ and\ \theta_0 < \theta_1$ such that

$R(\theta,\delta^*) \le \frac{I}{n} - \varepsilon \;\; for\ all\ \theta_0 < \theta < \theta_1$ (as in Theorem 4.3)

Let γ_T^* be the average Bayes risk of δ^* with respect to prior distribution $\tau \sim N(0,T^2)$ and let γ_T be the Bayes risk of the Bayes decision rule with respect to $N(0,T^2)$. Thus by exp. 3.11 for σ=1

$$\frac{\frac{1}{n}-\gamma_T^*}{\frac{1}{n}-\gamma_T} = \frac{\frac{1}{\sqrt{2\pi}T}\int_{-\infty}^{\infty}\left[\frac{1}{n}-R(\theta,\delta^*)\right]e^{\frac{-\theta^2}{2T^2}}d\theta}{\frac{1}{n}-\frac{T^2}{1+nT^2}}$$

$$\ge \frac{n(1+nT^2)\epsilon}{T\sqrt{2\pi}}\int_{\theta_0}^{\theta_1} e^{\frac{-\theta^2}{2T^2}}d\theta \qquad \ldots\ldots\ldots\ldots\ldots\ldots \ (4.15)$$

By Lebesgue dominated convergence theorem, as the integral

$e^{\frac{-\theta^2}{2T^2}} \to 1$ As T→ ∞, the integral converges to $(\theta_1-\theta_0)$ and the

R.H.S→ ∞ ⇒ $\frac{\frac{1}{n}-\gamma_T^*}{\frac{1}{n}-\gamma_T} \to \infty$ thus there exist T_0 such that, $\gamma_{T_0}^* < \gamma_{T_0}$, which contradicts the fact that γ_{T_0} is the Bayes risk for $N(0,T_0^2)$.

$R(\theta,\delta) = E(\delta-\theta)^2 = var_\theta(\delta) + b^2(\theta)$, where $b(\theta) = E_\theta(\delta) - \theta$

$$\ge b^2(\theta) + \frac{[1+b'(\theta)]^2}{nI(\theta)} \text{by F C R bound. } \ldots\ldots\ldots\ldots.. \ (4.16)$$

In the present case $\sigma^2 = 1, I(\theta) = 1$

Suppose now δ is any estimator satisfying

$$R(\theta,\delta) \le \frac{1}{n} For\ all\ \theta \ldots\ldots\ldots\ldots. \ (4.17)$$

and hence, $b^2(\theta) + \frac{[1+b'(\theta)]^2}{nI(\theta)} \le \frac{1}{n}\ \ for\ all\ \theta$ …………… (4.18)

We shall then show that (4.18) ⇒ $b(\theta) \equiv 0$ for all θ. i.e δ is unbiased.

1. Since $|b(\theta)| \leq \frac{1}{\sqrt{n}}$ the function b is bounded.
2. From the fact that $1 + b'^2(\theta) + 2b'(\theta) \leq 1 \Rightarrow b'(\theta) \leq 0$ so that b is non-increasing.
3. Next, there exists a sequence of $\theta_i \to \infty$ and such that $b'(\theta_i) \to 0$

For suppose that $b'(\theta)$ were bounded away from 0 as θ→ ∞, $say\ b'(\theta) \leq -\varepsilon\ for\ all\ \theta,$ then $b(\theta)$ can not be bounded

as θ→ ∞, which contradicts 1.

4. Analogically it is seen that there exist a square $\theta_i \to -\infty$ and such that $b'(\theta_i) \to 0$ .Thus $b(\theta) \to 0$ as $\theta \to \pm\infty$ with inequality (4.18). Thus $b(\theta) \equiv 0$ follows from 2. Since $b(\theta) \equiv 0 \Rightarrow b'(\theta) = 0\ for\ all\ \theta \Rightarrow (4.16)\ as\ R(\theta, \delta) \leq \frac{1}{n} For\ all\ \theta\ and\ hence\ R(\theta, \delta) \equiv \frac{1}{n}$

This proves that $\bar{X}$ is admissible and minimax. This is unique admissible and minimax estimator. Because if δ' is any other estimator such that $R(\theta, \delta') \equiv \frac{1}{n}$. Then let $\delta^* = \frac{1}{2}(\delta + \delta')$

$$R(\theta, \delta^*) < \frac{1}{2}[R(\theta, \delta) + R(\theta, \delta')] = R(\theta, \delta)$$

Which contradicts that δ is admissible. Thus δ=δ' with prob. 1.

3.6 Self-Assessment Exercise

1. Clearly differentiate between Bayes and Minimax Principles.
2. Discuss the concepts of Generalized Bayes Rule, Extended Bayes Rule and Limits of Bayes Rule along with their usefulness.

3.7 Summary

This unit explains the concepts of various structures of decision rules and hence enables the reader to make use of them in various decision-making situations. Section 3.3 discusses in detail about the Bayes and Minimax decision policies. Section 3.4, 3.5 and 3.6 cover the concepts of Generalized Bayes Rule, Extended Bayes Rule, and Limits of Bayes Rule.

3.8 Further Readings

1. Berger, J.O. (1993) Statistical Decision Theory and Bayesian Analysis, Springer Verlag.
2. Bernando, J.M. and Smith, A.F.M. (1994). Bayesian Theory, John Wiley and Sons.
3. Box, G.P. and Tiao, G.C. (1992). Bayesian Inference in Statistical Analysis, Addison-Wesley.
4. Robert, C.P. (1994). The Bayesian Choice: A Decision Theoretic Motivation, Springer.

UNIT-4: BAYESIAN INTERVAL ESTIMATION

Structure

4.1 Introduction

Estimation is the method of drawing conclusions regarding an unknown population parameter with the help of a sample from that population. Unlike point estimates, which are single-value estimates of a unknown population parameter, interval estimates are likely to contain the value of interest to a certain probability. Confidence intervals are the most well-known of the various forms of statistical intervals.

4.2 Objectives

After studying this unit, you should be able to

- Define the HPD intervals and credible sets.
- Obtain suitable techniques to derive the HPD regions.
- Solve questions in deriving HPD regions.

4.3 Bayesian Interval Estimation

In Bayesian approach, a credible interval is an interval in the domain of a posterior probability distribution, within which the value of the unknown parameter falls with certain probability.

In choosing a credible set for θ, it is usually described to try to minimize its size. To do this one should include in the set only those points with the largest posterior density i.e the most likely values of θ.

4.4 Credible Intervals

Definition: A $100(1-\alpha)\%$ credible set for θ is subset of Θ such that,

$$1-\alpha \leq P[C/x] = \int_C dF^{\pi/(\theta/x)}(\theta)$$

$$= \int_C \pi/(\theta/x)\, d\theta \quad \textit{for continuous case}$$

$$= \sum_{\theta \epsilon C} \pi/(\theta/x) \quad \textit{for discrete case}$$

Since the posterior distribution is an actual prob. distribution on Θ, one can speak of the probability that θ is C. this is in contrast to classical confidence procedures, which can only be interpreted in term of coverage probability that is the probability that the random variable X will be such the confidence set C(X) contains θ.

In choosing a credible set for θ, it is usually described to try to minimize its size. To do this one should include in the set only those points with the largest posterior density i.e the most likely values of θ.

Def: The $100(1-\alpha)\%$ HPD credible set (HPD region) for θ is the subset C of Θ of the form

$$C = \{\theta \epsilon \Theta: \pi(\theta/x) \geq K(\alpha)\}$$

$Where\ K(\alpha)$ Is the largest constant such that,

$$P[C/x] \geq 1-\alpha.$$

4.5 HPD Intervals

Exp: let $(X_1, \dots, X_n)$ be a random sample from N $(\theta, 1)$. Let the prior p.d.f of θ be N (μ, τ^2). Find the HDD regions for θ.

Solution: $f(\theta/x_1,\dots,x_n)=\frac{f(x_1,\dots,x_n/\theta)\pi(\theta)}{\int_{-\infty}^{\infty}f(x_1,\dots,x_n/\theta)\pi(\theta)d\theta}$

$$=\frac{exp-\frac{\Sigma(x_i-\bar{x})^2}{2}-\frac{n(\bar{x}-\theta)^2}{2}\exp-\frac{(\theta-\mu)^2}{2\tau^2}}{exp-\frac{\Sigma(x_i-\bar{x})^2}{2}\int_{-\infty}^{\infty}\exp-\frac{n(\bar{x}-\theta)^2}{2}\exp-\frac{(\theta-\mu)^2}{2\tau^2}d\theta} \quad =\frac{exp-\frac{n(\bar{x}-\theta)^2}{2}\exp-\frac{(\theta-\mu)^2}{2\tau^2}}{\int_{-\infty}^{\infty}\exp-\left[\frac{n(\bar{x}-\theta)^2}{2}\exp-\frac{(\theta-\mu)^2}{2\tau^2}\right]d\theta}$$

$$\int_{-\infty}^{\infty}\exp-\left[\frac{n(\bar{x}^2+\theta^2-2\bar{x}\theta)}{2}+\frac{(\theta^2+\mu^2+2\theta\mu)}{2\tau^2}\right]d\theta$$

$$=\exp\left(-\left(\frac{n\bar{x}^2}{2}+\frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty}\exp-\frac{1}{2}\left[\theta^2-2\theta\left(\bar{x}+\frac{\mu}{\tau^2}\right)+\frac{\mu^2}{\tau^2}\right]d\theta\ \right)$$

$$=\exp\left(-\left(\frac{n\bar{x}^2}{2}+\frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty}e^{-\frac{1}{2}}\left[\theta^2-2\theta\left(\bar{x}+\frac{\mu}{\tau^2}\right)+\frac{\mu^2}{\tau^2}\right]d\theta\right)$$

$$=\exp\left(-\left(\frac{n\bar{x}^2}{2}+\frac{\mu^2}{2\tau^2}\right)\int_{-\infty}^{\infty}e^{-\frac{1}{2}}\left[\theta^2-2\theta\left(\bar{x}+\frac{\mu}{\tau^2}\right)+\left(\bar{x}^2+\frac{\mu}{\tau^2}\right)-\left(\bar{x}^2+\frac{\mu}{\tau^2}\right)+\frac{\mu^2}{\tau^2}\right]d\theta\right)$$

$$=\exp\left(-\frac{n\bar{x}^2\tau^2+\mu^2}{2\tau^2}-\frac{1}{2}\left(\bar{x}+\frac{\mu}{\tau^2}\right)^2-\frac{\mu^2}{2\tau^2}\int_{-\infty}^{\infty}e^{-\frac{1}{2}}\left[\theta-\left(\bar{x}+\frac{\mu}{\tau^2}\right)\right]^2d\theta\ \right)$$

Let $\mu=0$

$$\int_{-\infty}^{\infty}\exp-\left[\frac{n\bar{x}^2+\theta^2-2\bar{x}\theta}{2}+\frac{\theta^2}{2\tau^2}\right]d\theta$$

$$=\exp\left(\frac{-n\bar{x}^2}{2}-\frac{\bar{x}^2}{2}\right)\int_{-\infty}^{\infty}\exp-\frac{1}{2}[\theta-\bar{x}]^2d\theta$$

$$=\sqrt{2\pi}\exp-\frac{1}{2}(-n\bar{x}^2+\bar{x}^2)$$

$$\therefore\quad \pi(\theta/x)=\frac{1}{\sqrt{2\pi}}\exp-\frac{(n\bar{x}-\theta)^2}{2}-\frac{\theta^2}{2\tau^2}+\frac{1}{2}(-n\bar{x}^2+\bar{x}^2)$$

$$=\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}\left[n\bar{x}^2+n\theta^2-2n\bar{x}\theta+\frac{\theta^2}{\tau^2}-n\bar{x}^2-\bar{x}^2\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2\tau^2} [n\theta^2\tau^2 - 2n\bar{x}\theta\tau^2 + \theta^2 - \bar{x}^2\tau^2]$$

$$= \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2\tau^2} [\theta^2(1 + n\tau^2) - 2n\bar{x}\theta\tau^2 - \bar{x}^2\tau^2]$$

$$\pi(\theta/x) = N(\mu(\bar{x}), P^{-1})$$

$$\mu(\bar{x}) = \frac{\tau^2\bar{x}}{\tau^2 + \frac{\sigma^2}{n}}, \quad P = \frac{n\tau^2 + \sigma^2}{\tau^2\sigma^2} \quad , \frac{1}{P} = \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}$$

4.6 Comparison with Classic Confidence Interval

In classical approach we consider that a parameter has one particular true value, and conduct an experiment whose resulting conclusion, irrespective of the true value of the parameter, will be correct with at least some minimum probability; while in Bayesian approach we say that the parameter's value is fixed but has been chosen from some probability distribution, called the prior probability distribution. This "prior" might be known or it might be an assumption drawn out of experience of the experimenter or otherwise. Clubbing this prior with the observed information Bayesians obtain the "posterior." Bayesian approaches can summarize their uncertainty by giving a range of values on the posterior probability distribution that includes 95% of the probability and this is called a "95% credibility interval.

4.7 Self-Assessment Exercise

1. Clearly differentiate between the Bayesian and classical interval estimation.

2. Discuss the concept of HPD intervals and its importance.

4.8 Summary

This unit aims in section 4.3, 4.4 and 4.5 at enabling the reader with the concept of interval estimation and to obtain the interval estimates from Bayesian point of view. And in section 4.6, the reader learns the difference between the classical and Bayesian approaches of interval estimations.

4.9 Further Readings

1 Gemerman, D and Lopes, H. F. (2006) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman Hall.

2 Lee, P.M. (1997) Bayesian Statistics: An Introduction, Arnold.

3 Leonard, T. and Hsu, J.S.J. (1999) Bayesian Methods, Cambridge University Press.

4 Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods, Springer Verlag.

U.P.RajarshiTandon Open University, Prayagraj

MScSTAT – 301N /MASTAT – 301N Decision Theory & Bayesian Analysis

Block: 2 Optimality and Decision Rules

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. RajarshiTandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P.RajarshiTandon Open University, Prayagraj

Course Preparation Committee

Dr. Pramendra Singh Pundir **Writer**
Department of Statistics
University of Allahabad, Prayagraj

Prof. G. S. Pandey (Rtd.) **Editor**
Department of Statistics
University of Allahabad, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. RajarshiTandon Open University, Prayagraj

MScSTAT – 301N/ MASTAT – 301N DECISION THEORY & BAYESIAN ANALYSIS

©UPRTOU

First Edition: ***July 2023***

ISBN :

Printed and Published by Col. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2023.

Printed By:

Block & Unit Introduction

The present block of this SLM has four units.

The ***Block - 2 – Optimality of Decision Rules*** is the second block with four units, which impasses about the different rules.

In ***Unit – 5 – Admissibility and Completeness*** is discussed with respect to Bayes rule and prior distribution minimal complete class.

In ***Unit – 6 – Minimaxity and Multiple decision Problem*** has been introduced, along with complete class theorem and admissibility rules. Equalizer rules have been discussed and maximin and minimax strategies have been explained.

Unit – 7 – Bayesian Decision Theory dealt with theorem on optimal Bayes decision function, Relationship of Bayes and minimax decision rules and least favourable distributions.

Unit – 8 – Bayesian Inference dealt with Bayesian sufficiency, On informative Priors, Improper prior densities

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT-5: ADMISSIBILITY AND COMPLETENESS

Structure

5.1 Introduction

Admissibility refers to a set of rules for making a decision such that no other rule exists which is always better than the defined rules.

5.2 Objectives

After studying this unit, you should be able to

- Define admissibility of a set of rules.
- Check for admissibility with respect to Bayes' rules.
- Define completeness and minimal complete class.

5.3 Admissibility

<u>Theorem 4.2</u>: Assume that $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ and a Bayes ruleδ_0 w.r.to the prior distribution $(p_1, p_2, \dots, p_k)$ exists. If $p_j > 0$ for j=1, 2, …, k, then δ_0 is admissible.

Proof: Suppose that δ_0 is inadmissible, then there exist a $\delta' \varepsilon D^*$

which is better than δ_0. That is,

$$R(\theta_j, \delta') \leq R(\theta_j, \delta_0) \qquad \text{for all j}$$

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

Because, all p_j are positive

$$\sum R(\theta_j, \delta') p_j < \sum p_j R(\theta_j, \delta_0)$$

The strict inequality showing that δ_0 is not Bayes w.r.to$(p_1, p_2, \dots, p_k)$. This is a contradiction.

The following counter example shows that δ_0 is not necessarily admissible if the hypothesis $p_j > 0$ for j=1, 2,…,k is violated.

Ex 4.1: let $\Theta = \{\theta_1, \theta_2\}$, $L(\theta, a)$ as follows:

		a_1	a_2	a_3	a_4
$L(\theta, a)$	θ_1	1	1	2	2
	θ_2	0	1	0	1

$d(0) = a_1$, $d(0) = a_2$, $d(0) = a_3$, $d(0) = a_4$

$$R(\theta_1, a_1) = 1, \; R(\theta_2, a_1) = 0, \dots\dots, R(\theta_1, a_4) = 2, R(\theta_2, a_4) = 1$$

$$R(\theta_1, \delta) = \sum_{i=1}^{4} \alpha_i R(\theta_1, a_1) S = \{R(\theta_1, \delta), R(\theta_2, \delta): \qquad \delta \in D^*\} R(\theta_2, \delta)$$

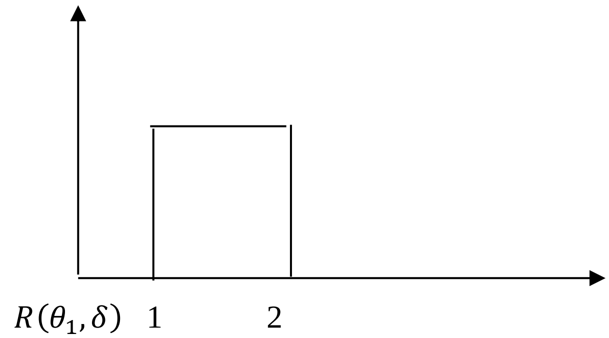
$$= \{(y_1, y_2): \quad 1 \leq y_1 \leq 2; 0 \leq y_2 \leq 1\}$$

Bayes rule w.r.to (1,0)

Let the prior distribution, $p_1 = 1, p_2 = 0$

$\sum_{i=1}^{4} p_i R(\theta_i, \delta) = R(\theta_1, \delta) = y_1$

S



Thus, any decision rule that minimizes $\sum p_i R(\theta_i, \delta)$ and that achieved the minimum value =1=y_1 will be a Bayes rule w.r.to prior (1, 0).

Thus the rule $R(\theta_1, \delta_0) = R(\theta_2, \delta_0) = 1$ is Bayes w.r.to (1, 0).that a_2 and a_1 are Bayes rules w.r.to (1,0). But a_2 is not admissible since

$R(\theta_1, a_2) \leq R(\theta_2, a_1)$and $R(\theta_2, a_2) > R(\theta_2, a_1)$.

<u>Def 4.5</u>: A point θ_0 in E_1(one dimensional Euclidian space) is said to be in support of a distribution τ on the real line if for ∀ε> 0 the interval $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ has positive probability,

$$\tau(\theta_0 - \varepsilon, \theta_0 + \varepsilon) > 0$$

<u>Theorem 4.3</u>: let $\Theta \varepsilon E_1$ and assume that $R(\theta, \delta)$ is a continuous function of θ for all $\delta \varepsilon D^*$. If δ_0 is a Bayes rule w.r.to a probability distribution τ on the real line, for which $\Upsilon(\tau, \delta_0)$ is finite and if the support of τ is the whole real line, then δ_0 is admissible.

<u>Proof</u>: As before, assume that δ_0 is not admissible. Then, there exists a $\delta' \varepsilon D^*$for which

$$R(\theta, \delta') \leq R(\theta, \delta_0) \qquad \text{for all } \theta.$$

$$R(\theta_0, \delta') < R(\theta_0, \delta_0) \qquad \text{for some } \theta_0 \varepsilon E_1.$$

Since $R(\theta, \delta)$ is continuous in θ for all δ. Let

$\eta = R(\theta_0, \delta_0) - R(\theta, \delta')$…………………. (4.1)

For $|\theta - \theta_0| < \varepsilon \quad , \varepsilon > 0$

$|R(\theta, \delta) - R(\theta_0, \delta)| < \frac{\eta}{4}$ Whenever $|\theta - \theta_0| < \varepsilon \ for\ all\ \delta \varepsilon D^*$

Or $\quad -\frac{\eta}{4} \leq R(\theta, \delta) - R(\theta_0, \delta) \leq \frac{\eta}{4} |\theta - \theta_0| < \varepsilon$ ……………. (4.2)

Or $\quad R(\theta, \delta) \leq R(\theta_0, \delta) + \frac{\eta}{4}$

$$R(\theta, \delta') \leq R(\theta_0, \delta') + \frac{\eta}{4} \quad \text{for all } |\theta - \theta_0| < \varepsilon$$

$$= R(\theta, \delta_0) - R(\theta, \delta_0) + R(\theta_0, \delta') + \frac{\eta}{4}$$

$$= R(\theta, \delta_0) - [R(\theta, \delta_0) - R(\theta_0, \delta_0) + R(\theta_0, \delta_0) - R(\theta_0, \delta')] + \frac{\eta}{4}$$

$$= R(\theta, \delta_0) - [R(\theta, \delta_0) - R(\theta_0, \delta_0)] - [R(\theta_0, \delta_0) - R(\theta_0, \delta')] + \frac{\eta}{4}$$

$$\leq R(\theta, \delta_0) + \frac{\eta}{4} - \eta + \frac{\eta}{4} = R(\theta, \delta_0) - \frac{\eta}{2}$$

Thus, $R(\theta, \delta') \leq R(\theta, \delta_0) - \frac{\eta}{2}$ whenever $|\theta - \theta_0| < \varepsilon$

Letting T denote the r.v. whose d.f is τ

$\Upsilon(\tau, \delta_0) - \Upsilon(\tau, \delta') = \mathrm{E}\, R(T, \delta_0) - \mathrm{E}\, R(T, \delta')$

$$= \mathrm{E}\, [R(T, \delta_0) - R(T, \delta')] = \int R(t, \delta_0) - R(t, \delta')\, d\tau$$

$$= \int_{|\theta - \theta_0| < \varepsilon} [R(t, \delta_0) - R(t, \delta')] d\tau + \int_{|\theta - \theta_0| \geq \varepsilon} [R(t, \delta_0) - R(t, \delta')] d\tau$$

$$\geq \int_{|\theta - \theta_0| < \varepsilon} [R(t, \delta_0) - R(t, \delta')] d\tau \ \geq \frac{\eta}{2}\, \tau(\theta - \varepsilon, \theta + \varepsilon)$$

That is δ_0 is not Bayes rule, which is a contradiction. #

<u>Def 4.6</u>: A set, S, k- dimensional Euclidian space, E_k , is said to be *bounded from below* if there exists a finite number M, such that for every $y = (y_1, y_2, \dots, y_k)\varepsilon\, S y_j > -M\ for\ j = 1, \dots, k$ ……….(4.3)

Thus, a set S is bounded from below if for each fixed j,1≤j≤k the coordinate y_j is bounded below as y ranges through S.

<u>Def 4.7</u>: Let x be a point in E_k. The *lower quant ant at* x, denoted by Q_x is defined as the set

$Q_x = \{y\varepsilon E_k: y_j \le x_j\ for\ j = 1, \dots, k\}$……………… (4.4)

$Thus\ Q_x$is a set of risk points as good as x and $Q_x - \{x\}$ is the set of risk points better than x. $\bar{S}$is the smallest closed set containing S.

<u>Def 4.8</u>: A point x is said to be a *lower boundary point* of a convex set S⊂E_k if$Q_x \cap \bar{S} = \{x\}$. The set of lower boundary points of a convex set is defined by λ(S).

<u>Def 4.9</u>: A convex set S⊂E_k is said to be *closed from below* if λ(S) ⊂S.

<u>Theorem 4.4</u>: Suppose that $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ and the risk set S is bounded from below and closed from below. For every prior distribution $(p_1, p_2, \dots, p_k)$ for which$p_j > 0$ for all j=1,…,k, a Bayes rule w.r.t. $(p_1, p_2, \dots, p_k)$ exists.

<u>Proof</u>: Let $(p_1, p_2, \dots, p_k)$ 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 $= \sum p_j\, y_j$, where $y = (y_1, y_2, \dots, y_k)\varepsilon\ S$

$$B = \left\{ b = \sum_j p_j y_j\ for\ some\ y\varepsilon\ S \right\}$$

Because S is bounded from below, so is B; let b_0 be the g. l. b. of B. in a sequence of points $y^{(n)}\varepsilon\ S$ 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 $y^0 \varepsilon$ λ(S). Since y^0 is a limit point of points of S, $y^0 \varepsilon\, \overline{S}$ and $\{y^0\} \subset Q_{y^0} \cap \bar{S}$. Further more$Q_{y^0} \cap \bar{S} \subset \{y^0\}$, for if y' is any point of Q_{y^0} other thany^0 itself, $\sum p_j\, {y_j}' < b_0$ so that if

$y' \varepsilon\, \overline{S}$ There would exist point y of S for which$\sum p_j\, y_j < b_0$. This contradicts the assumption that b_0 is the lower bound of B. Thus

$Q_{y^0} \cap \bar{S} = \{y^0\}$,implying that $y^0 \varepsilon$ λ(S).

<u>Theorem 4.5</u>: Suppose that $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ and the risk set S is bounded from below and closed from below, the class of decision rules, $D_0 = \{\delta \varepsilon D^* : R(\theta_1, \delta), \dots, R(\theta_k, \delta) \varepsilon$ λ(S)} ………… (4.5)

Then,D_0a minimal complete class.

<u>Proof</u>: 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 non empty, convex,

[Since closer of convex set is convex and the intersection of two convex sets is convex.] and bounded below. Thus $\lambda(S_1)$ is non empty (by theorem 4.4). Let $y\varepsilon\ \lambda(S_1)$; then $\{y\} = Q_y \cap \overline{S_1}$ further y ε $Q_x$because $y\varepsilon \overline{S_1} = \overline{Q_x \cap \bar{S}} \subset \overline{Q_x} = Q_x$. Finally $y\ \varepsilon\ \lambda(S)$ because

$\{y\} = Q_y \cap \overline{S_1} = Q_y \cap \overline{Q_x \cap \bar{S}} = Q_y \cap Q_x \cap \bar{S} = Q_y \cap \bar{S}$.

Thus, because S is closed from below, there exists a $\delta_0 \varepsilon D_0$ for which

$y = \{R(\theta_1, \delta_0), \dots, R(\theta_k, \delta_0)\}$,and which is better than δ since,

$y\ \varepsilon Q_x - \{x\}$. This proves D_0 is complete.

Since every rule in D_0 is admissible. Hence no proper subset of D_0 should be complete. Because, every complete class must contain all admissible rules, thus D_0 is minimal complete.

5.4 Completeness

After the all discussion, now we are ready to learn the following definitions and theorems:
Definition: A class C of decision rules is said to be complete if, for any decision rule δ not in C, there is a decision rule δ' in C, which does not have less risk than δ.
Definition: A class C of decision rules is said to be minimal complete if C is complete and if no proper subset of C is complete.

5.4 Minimal Complete Class

Definition: A class C of decision rules is said to be complete if, for any decision rule δ not in C, there is a decision rule δ' in C, which does not have less risk than δ.
Definition: A class C of decision rules is said to be minimal complete if C is complete and if no proper subset of C is complete.

5.6 Separating and Supporting Hyper Plane Theorems

Lemma 4.2: If S is closed convex set of E_k and 0∉S, then there exists a vector P∈E_k such that $P^T x > 0$ for all x ε S.

Proof: For every real number $\alpha > 0$ let B_α is the sphere of radius α centered at origin. $B_\alpha = \{x \varepsilon E_k : x^T x \leq \alpha^2\}$. Let A be the set of all real $\alpha > 0$ for which B_α intersects S, $A = \{\alpha : B_\alpha \cap S \neq \Phi\}$. Because the Lemma is trivial if S is empty, we consider that S is non empty. Hence A is non empty. Let a = g l b of A. a is finite because A is non empty and positive because S is closed and 0∉S.

1. $B_\alpha \cap S$ is non empty. As α→ a from above $B_\alpha \cap S$ is a decreasing intersection of non empty compact sets whose limit $B_\alpha \cap S$ is therefore non empty.
2. For all $x \; \varepsilon \; S, P^T(x - P) \geq 0$. Let f(β) denote the square of the distance from the origin to the part $\beta x + \overline{1 - \beta} P$ for a fixed $x \; \varepsilon \; S, x \neq P$

$$f(\beta) = \left(\beta x + \overline{1-\beta}P\right)^T \left(\beta x + \overline{1-\beta}P\right)$$

$$= \beta^2 (x-P)^T (x-P) + 2\beta P^T (x-P) + P^T P \text{ ....} \quad (4.6)$$

(4.6) $will\ beminimum\ if\ \beta = \beta_0$ where,

$$\beta_0 = -\frac{P^T(x-P)}{(x-P)^T(x-P)} \text{ ....} \quad (4.7)$$

Because,$f(1) = x^T x \geq P^T P = f(0)$, it is clear that $\beta_0 < 0$, further since $\beta x + \overline{1-\beta}P \in$ S ,where $0 < \beta < 1$ from the convexity of S. it is clear that $\beta_0$can not be $0 < \beta_0 < 1$ without contradicting the fact that no point of S is closer to the origin than P. $Hence\ \beta_0 \leq 0. Or$ equivalently,

$$P^T(x-P) \geq 0 \Rightarrow P^T x \geq P^T P > 0 \quad for\ all\ x\varepsilon S.$$

#

<u>Lemma 4.3</u>: If S is convex subset of E_k, A is open subset of E_k, and $A \subset \bar{S},\ \ then\ A \subset S.$

<u>Theorem 4.6</u>:(Supporting Hyper Plane Theorem): If S is closed convex sub set of E_k and x_0 is not an interior point of S (i.e. either$x_0 \notin S$

orx_0 is a boundary point of S), then there exists a vector$P\varepsilon E_k$, P$\neq$ 0

Such that $P^T x \geq P^T x_0$ for all x ε S.

<u>Proof</u>: Becausex_0 is not an interior point of S,x_0 is not an interior point of $\bar{S}$ by Lemma (4.3). Hence there is a sequence $y_n \notin \bar{S}$ for which$y_n \to x_0$. We shall translate the origin toy_n successively and applying Lemma (4.2). Let

$$S_n = \{Z : Z = x - y_n, x\varepsilon S\}$$

Then $\bar{S}_n$ closed convex set, and $0 \notin \bar{S}_n$. From Lemma (4.2) there exists a vector $P_n \varepsilon E_k$ such that $P_n^T Z > 0$ for all $Z\varepsilon \bar{S}_n$ or$P_n^T(x - y_n) > 0$

$For\ all\ x\varepsilon\bar{S}$. Let $q_n = \frac{P_n}{\sqrt{P_n^T P_n}}$. Then $q_n^T q_n = 1$ because unit sphere in E_k is compact, there exists a limit point P of $the\, q_n and$ a subsequence $q_{n'} \to P$. Hence $q_{n'}^T(x - y_n) \to P^T(x - x_0)$, but $q_{n'}^T(x - y_n) > 0$ for all x ε S ⇒ $P^T(x - x_0) \geq 0$ for all x ε S as was to be proved.

<u>Theorem 4.7</u>:(Separating Hyper Plane Theorem): Let $S_1$and S_2 be disjoint convex subsets of E_k then there exists a vector $P \neq 0$ such that $P^T y \leq P^T x$ for all x εS_1 and y ε$S_{2.}$

<u>Proof</u>: Let $S = \{Z: Z = x - y\ for\ some$ x εS_1 and y ε$S_{2.}\}$

1. S is convex. Let Z_1, Z_2 elements of S and let 0< $\beta < 1$. We are to show that$\beta Z_1 + \overline{1 - \beta} Z_2 \varepsilon\, S$. Let x_1,$x_2 \varepsilon\, S_1$, $y_1, y_2 \varepsilon\, S_2$ such that

   $Z_1 = x_1 - y_1, Z_2 = x_2 - y_2\ \varepsilon\, S$ Then,

   $$\beta Z_1 + \overline{1 - \beta} Z_2 = \beta(x_1 - y_1) + \overline{1 - \beta}(x_2 - y_2\,)$$

   $$= \left(\beta x_1 + \overline{1 - \beta} x_2\right) - \left(\beta y_1 + \overline{1 - \beta} y_2\,\right)\varepsilon\, S \quad as$$

   $$\beta x_1 + \overline{1 - \beta} x_2 \varepsilon\, S_1\ \ , \beta y_1 + \overline{1 - \beta} y_2\ \varepsilon\, S_2 \Rightarrow S\ is\ convex.$$

2. $0 \notin S$ For if $0 \in S$, there could be point x εS_1,y ε$S_{2.}$ such that (x-y)=0 ⇒ x=y contradicts that S_1 and S_2 are disjoint.
3. From Theorem (4.6) there exists a vector $P \neq 0$ such that $P^T Z \geq 0$ for all Z ε S. Thus$P^T(x - y) \geq 0\ for\ all$ x εS_1,y ε$S_{2.}$ completing the proof.

<u>Lemma 4.4</u>: If S is a convex sub set of E_k and Z is a k-dimensional random vector for which E (Z) exists and is finite, then EZ∈S.

<u>Proof</u>: Let Y=Z-EZ and let S’ be the translation of S by E Z, i.e$S' = \{Y: Y = Z - EZ\ \ for\ all\ Z\varepsilon S\}$. Thus S’ is convex $P[Y\varepsilon S'] = 1$ and EY=0. We will show that$0 \in S'$. We prove by induction method. The Lemma is trivially true for k=0 in which case Y is degenerate at zero. Now suppose the Lemma is true for k-1. We are to show that Lemma is true for k≥ 1.

Suppose $0 \notin S'$ then by Theorem (4.6) there exists a vector $P \neq 0$ such that $P^T Y \geq 0$ for all Y ε S'. Let U=$P^T Y$. The r.v. U has expectation 0, and $P[U \geq 0] = 1 \Rightarrow P[U = 0] = 1$, then with probability one Y lies in the hyper plane$P^T Y = 0$. Let

$S'' = S' \cap \{y \colon P^T Y \ = 0\}$ Then S'' is convex subset of (k-1) dimensional Euclidian space for which $P[Y \varepsilon S''] = 1 \ \ and \ \ EY = 0$

$By\ the\ induction,$ $0 \varepsilon S''$. Since $S'' \subset S' \Rightarrow 0 \varepsilon S'$ which is contradiction of the assumption $0 \notin$ S'.

#

<u>Corollary</u>: S is a convex hull of S_0.

<u>Lemma 4.5</u>:(Jensen's Inequality): Let f(x) be a convex real-valued function defined on a non-empty convex subsets of E_k and let Z be a k-dimensional random-vector with finite expectation E Z for which $P[Z \in S] = 1$. Then E(Z)∈S and $f[E(Z)] \leq E[f(Z)]$ .... (4.8)

<u>Proof</u>: for k=1, the point $(EZ, f\ (EZ))$ is on the boundary of the convex set S_1.

$$S_1 = \left\{ \begin{matrix} (Z_1, Z_2, \dots, Z_{k+1})^T\ for\ some\ x \varepsilon S, x^T = (Z_1, Z_2, \dots, Z_{k+1}) \\ and\ f(x) \leq Z_{k+1} \end{matrix} \right\}. \ (4.9)$$

Hence there exists a supporting hyper plane (straight line) at

$\big(EZ, f\ (EZ)\big)$. Call this $y = m\,x + \ c$

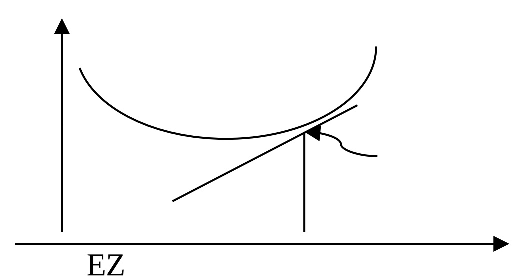
Because $(EZ, f\ (EZ))$ is on this line. It may be written as,

$Y = f(EZ) + m(x - EZ)$ And because this line is never above the curve y=f(x) we have, f(x)

$f(x) \geq f(EZ) + m(x - EZ)\ for\ all\ x.$

$f(Z) \geq f(EZ) + m(Z - EZ) \quad for\ Z \varepsilon\ S.$[EZ, f (EZ)]

$E(f(Z)) \geq f(EZ)$



Thus, theorem is true for k=1. Suppose theorem is true for k-1, we prove for k≥ 1.

Since EZ ε S, the point $(EZ, f\ (EZ))$ is boundary point of the convex set S_1 defined (4.9) hence by supporting hyper plane theorem, there exists $a(k+1)$-dimensional vector $P \neq 0$ such that,

$$P^T Z \geq P^T\big(EZ, f\ (EZ)\big) \quad or$$

$\sum_{j=1}^{k+1} p_j\ z_j \geq \sum_{j=1}^{k} p_j\ Ez_j + p_{k+1} f(Ez)\ \ for\ all (Z_1, \dots, Z_k)^T \varepsilon S_1$. (4.10)

We note that;p_{k+1} can not be negative, for letting $Z_{k+1} \to \infty$ the inequality (4.10) will not be satisfied. Replacing Z_{k+1}

$with\ f(Z), Z = (Z_1, \dots, Z_k) \varepsilon S$and Z with random vector Z.

$p_{k+1} f(EZ) \leq \sum_{j=1}^{k+1} p_j\ (z_j - Ez_j) + p_{k+1} f(Z)$ …………… (4.11)

If $p_{k+1} > 0$ taking the expectation.

$p_{k+1} f(EZ) \leq p_{k+1} Ef(Z) \Rightarrow f[E(Z)] \leq E[f(Z)]$

If $p_{k+1} = 0$ (4.11) ⇒ the random vector

$U = \sum p_j (z_j - Ez_j) = P^T(z - Ez)$is non-negative and EU=0⇒P[U=0]=1 that gives all its mass to the (k-1) dimensional convex set $S' = S \cap \{Z : \sum p_j (z_j - Ez_j) = 0\}$ by induction method, theorem is proved.

Theorem 4.8: Let â be a convex subset of E_k and let L (θ, a) be a convex function of a ε â for all θ ε Θ there exist a ε> 0 and a c such that$L(\theta', a) \geq \varepsilon|a| + c$, then for every P ε â*, there exist $an\ a_0 \varepsilon$â such that $L(\theta, a_0) \leq L(\theta, P)$ for all θ ε Θ.

Proof: P ε â* and Z be a random vector with values in â when distribution is given by P. then EZ infinite since,

$\varepsilon E|Z| + c \leq EL(\theta', Z) = L(\theta', P) < \infty$ By definition of â*.

$$L(\theta, P) = EL(\theta, Z) \geq L(\theta, EZ) = L(\theta, a_0) Where, a_0 = EZ\varepsilon\hat{a}.$$

Remark: If the loss is convex, we can always concern with non-randomized decision rules. The non-randomized decision rules form a complete class.

Exp 4.2: Θ = â = [0,1], â is convex set.

$L(\theta, a) = (\theta - a)^2$ is convex loss function.

X has b= (2, θ)

$$P_\theta[X = x] = \binom{2}{x}\theta^x(1-\theta)^{2-x} \qquad x = 0,1,2$$

$$d_1(x) = \frac{x}{2} \quad d_2(x) = \frac{1}{2} \qquad for\ all\ x = 0,1,2$$

$$P[Z = d_1] = \frac{1}{2} \qquad P[Z = d_2] = \frac{1}{2}$$

$$E[Z] = \frac{d_1 + d_2}{2} = \frac{x+1}{4} = d$$

$$R(\theta, d) = EL\big(\theta. d(x)\big) = E(\theta - \frac{x+1}{4})^2$$

$$= \theta^2 + E(\frac{x+1}{4})^2 - 2\theta E(\frac{x+1}{4})$$

$$= \theta^2 + \frac{1}{16}[Ex^2 + 1 + 2Ex] - \frac{\theta}{2}(E(x) + 1)$$

$$= \theta^2 + \frac{1}{16}[2\theta(1-\theta) + 4\theta^2 + 1 + 2.2\theta] - \frac{\theta(2\theta+1)}{2}$$

$$= \frac{16\theta^2 + [2\theta - 2\theta^2 + 4\theta^2 + 1 + 4\theta] - 16\theta^2 - 8\theta}{16} = \frac{[2\theta^2 - 2\theta + 1]}{16}$$

Let d_0 be a randomized decision rule choosing d_1 with prob. $\frac{1}{2}$ and

d_2 with prob. $\frac{1}{2}$

$$R(\theta, d_0) = \frac{1}{2}[R(\theta, d_1) + R(\theta, d_2)]$$

$$= \frac{1}{2}\left[\frac{1}{2}\theta(1-\theta) + \frac{1}{4}(4\theta^2 - 4\theta + 1)\right] = \frac{1}{8}(2\theta^2 - 2\theta + 1)$$

Obvious, $R(\theta, d) \leq R(\theta, d_0)$ $\ as$

$$\frac{[2\theta^2-2\theta+1]}{16} \leq \frac{(2\theta^2-2\theta+1)}{8}$$

$$2\theta^2-2\theta+1 \geq 0 \qquad 1-2\theta(1-\theta) \geq 0$$

as the maximum value of $,\theta(1-\theta)=1/4$. Thus, the inequality is always true.

5.7 Self-Assessment Exercise

1. If g is a continuous and concave function on the interval I and X is a r.v. whose values are in I, with certainty, then E[g(X)] ≤ g[E(X)], provided expectations exist.

2. State and prove supporting and separating hyper plane theorems along with their uses.

5.8 Summary

Section 5.3 discusses the about the concept of admissibility. Concepts of completeness and minimal complete class and related results have been covered in sections 5.4 and 5.5. Separating and Supporting Hyperplane Theorems and some others important results and their derivations are given in section 5.6.

5.9 Further Readings

1. Berger, J.O. (1993) Statistical Decision Theory and Bayesian Analysis, Springer Verlag.
2. Bernando, J.M. and Smith, A.F.M. (1994). Bayesian Theory, John Wiley and Sons.
3. Luenberger, David G. (1969). Optimization by Vector Space Methods. New York: John Wiley & Sons. p. 133.

UNIT-6: MINIMAXITY AND MULTIPLE DECISION PROBLEM

Structure

6.1 Introduction

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. Minimax theorems state that a wide variety of two-person zero-sum games have values and are strictly determined. A multiple decision problem is a problem in which only a finite set of actions (more than 2), is available.

6.2 Objectives

After studying this unit, you should be able to

- Define the minimax theorem.
- State the complete class theorem.

- Define multiple decision problems.
- State the continuous form of Bayes' theorem.

6.3 Minimax Theorem

Minimax theorem

As discussed in earlier sections, now we learn the concept of minimax theorems, which state that a wide variety of two-person zero-sum games have values and are strictly determined. In particular, if parametric space is finite (and certain technical conditions hold), then the game has a value and is strictly determined i.e. these theorem state that the game has a value and that minimax rules exist.

6.4 Complete Class Theorem

Theorem 4.9: (converse of theorem 4.2): If δ is admissible and Θ is finite, then δ is Bayes w.r.to some prior distribution τ.

Proof: If δ is admissible, then $Q_x \cap S = \{x\}$ where $x = \{R(\theta_1, \delta), \dots \dots \dots, R(\theta_k, \delta)\}$ as $S \subset \bar{S} \Rightarrow Q_x \cap S \subset Q_x \cap \bar{S} = \{x\}$. And x ε S. thus, because Q_x -{x} and S are disjoint convex sets, there exists a vector P≠ 0 such that $P^T y \leq P^T z$ for all $y \varepsilon Q_x$ -{x}, and z ε S. If some coordinate p_j of vector P were negative then by taking y so that y_j sufficiently negative, we would have $P^T y < P^T x$. Hence $p_j \geq 0$ for all j. we may normalize P so that $\sum p_j = 1$. Because P is now a probability

Distribution over Θ and $\sum p_j \, R\left(\theta_j, \delta\right) \leq P^T Z$ for all Z ε S, δ is a Bayes rule w.r.to P.

Theorem 4.10:(Complete Class Theorem): 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.

Exp 4.3: $\Theta = \{\theta_1, \theta_2\}$ â = [0,1]

$$L\,(\theta_1, a) = a^2, \qquad L\,(\theta_2, a) = 1 - a$$

(Note that loss function is convex in a, for each θ)

$$P_{\theta_1}\{H\} = \frac{1}{3} \qquad P_{\theta_2}\{H\} = \frac{2}{3}$$

1. Represent the class D rules as a subset of the plane.
2. Find the class of all non-randomized rules.
3. Find minimax Bayes rules.

Solution: $D = \{d: \mathfrak{x} \to [0,1]\}\ where\ \mathfrak{x} = \{H, T\}$

Let $d(H) = x$, $d(T) = y$ with the interpretation that we estimate θ to be x when H is observed and y when T is observed.

$$D = \{(x,y): 0 \le x \le 1, 0 \le y \le 1\}$$

This is a square in the plane (x, y).

$$R(\theta_1, d) = EL(\theta_1, (x,y))$$

$$= L(\theta_1, x)P\left[{}^{H}/_{\theta_1}\right] + L(\theta_1, y)P\left[{}^{T}/_{\theta_1}\right]$$

$$= x^2\tfrac{1}{3} + y^2\tfrac{2}{3} = \tfrac{1}{3}(x^2 + 2y^2) \text{ ………………….. (4.12)}$$

$$R(\theta_2, d) = EL(\theta_2, (x,y))$$

$$= L(\theta_2, x)P\left[{}^{H}/_{\theta_2}\right] + L(\theta_2, y)P\left[{}^{T}/_{\theta_2}\right]$$

$$= (1-x)\tfrac{2}{3} + (1-y)\tfrac{1}{3} = \tfrac{1}{3}(3 - 2x - y) \text{ ……………….. (4.13)}$$

Let (p) and (1-p) be the probability distribution $\Theta = \{\theta_1, \theta_2\}$ i.e choosing θ_1 with prob. (p) and choosing θ_2 with prob. (1-p).

$$R(\tau, (x,y)) = ER(\theta, (x,y))$$

$$= pR(\theta_1, (x,y)) + 1 - pR(\theta_2, (x,y))$$

$$= \frac{p}{3}(x^2 + 2y^2) + \frac{1-p}{3}(3 - 2x - y)$$

$$= \frac{p}{3}(x^2 + 2y^2 + 2x + y - 3) + \frac{1}{3}(3 - 2x - y) \text{ .... } (4.14)$$

Set of Bayes rules which minimizes (4.14) will be obtained as,

$(2x + 2)\frac{p}{3} - \frac{2}{3} = 0 \Rightarrow x = \frac{1-p}{p}$&

$(4y + 1)\frac{p}{3} - \frac{1}{3} = 0 \Rightarrow y = \frac{1}{4}\left(\frac{1-p}{p}\right)$

Then the set of Bayes rules are,

$$B = \left\{\left(\alpha, \frac{\alpha}{4}\right) : 0 \leq \alpha \leq 1\right\} \subset D.$$

Now to find minimax Bayes rule, we should have (4.12) = (4.13) for $\left(\alpha, \frac{\alpha}{4}\right) \in B \Rightarrow$

$\frac{1}{3}(\alpha^2 + \frac{2\alpha^2}{16}) = \frac{1}{3}(3 - 2\alpha - \frac{\alpha}{4})$

$\frac{9\alpha^2}{18} = 3 - 2\alpha - \frac{\alpha}{4} \Rightarrow 9\alpha^2 + 18\alpha - 24 = 0 \Rightarrow 3\alpha^2 + 6\alpha - 2 = 0$

$$\alpha = \frac{-6 \pm \sqrt{36+96}}{6} = -1 \pm \frac{5.74}{3} = 0.91, \qquad as\ \alpha \geq 0$$

$\frac{1-p}{p} = 0.91 \Rightarrow p = 0.52\ (approx.)$

Hence (0.52, 0.48) is prior distribution function (0.91, 0.23) is Bayes rule and since for this (x, y) risk is constant have (0.91, 0.23) is minimax Bayes rule.

Example: 4.4

Admissibility of $\bar{X}$ for estimating normal mean:

First proof: (the limiting Bayes method): Suppose $\bar{X}$ is not admissible, and without loss of generality we may assume σ=1. Then there exists δ^* such that

$$\left.\begin{aligned} R(\theta, \delta^*) &\leq \frac{I}{n}\ \ for\ all\ \theta \\ &< \frac{I}{n}\ for\ some\ \theta \end{aligned}\right\}\ (under\ the\ square\ error\ loss\ function)$$

R (θ, δ) is a continuous function of θ for every δ, so that there exist

$\varepsilon > 0 \; and \; \theta_0 < \theta_1$ such that

$R(\theta, \delta^*) \leq \frac{I}{n} - \varepsilon \;\; for \; all \; \theta_0 < \theta < \theta_1$ (as in Theorem 4.3)

Let γ_T^* be the average Bayes risk of δ^* with respect to prior distribution $\tau \sim N(0, T^2)$ and let γ_T be the Bayes risk of the Bayes decision rule with respect to $N(0, T^2)$. Thus by exp. 3.11 for σ=1

$$\frac{\frac{1}{n} - \gamma_T^*}{\frac{1}{n} - \gamma_T} = \frac{\frac{1}{\sqrt{2\pi}T} \int_{-\infty}^{\infty} \left[\frac{1}{n} - R(\theta, \delta^*)\right] e^{\frac{-\theta^2}{2T^2}} d\theta}{\frac{1}{n} - \frac{T^2}{1+nT^2}}$$

$$\geq \frac{n(1+nT^2)\epsilon}{T\sqrt{2\pi}} \int_{\theta_0}^{\theta_1} e^{\frac{-\theta^2}{2T^2}} d\theta \qquad \ldots\ldots\ldots\ldots\ldots\ldots \; (4.15)$$

By Lebesgue dominated convergence theorem, as the integral

$e^{\frac{-\theta^2}{2T^2}} \to 1$ As T→ ∞, the integral converges to $(\theta_1 - \theta_0)$ and the

R.H.S→ ∞ ⇒ $\frac{\frac{1}{n} - \gamma_T^*}{\frac{1}{n} - \gamma_T} \to \infty$ thus there exist T_0 such that, $\gamma_{T_0}^* < \gamma_{T_0}$, which contradicts the fact that γ_{T_0} is the Bayes risk for $N(0, T_0^2)$.

<u>Second proof</u>: (the information inequality method):

$R(\theta, \delta) = E(\delta - \theta)^2 = var_\theta(\delta) + b^2(\theta)$, where $b(\theta) = E_\theta(\delta) - \theta$

$\geq b^2(\theta) + \frac{[1+b'(\theta)]^2}{nI(\theta)}$ by F C R bound. ………….. (4.16)

In the present case $\sigma^2 = 1, I(\theta) = 1$

Suppose now δ is any estimator satisfying

$R(\theta, \delta) \leq \frac{1}{n} For \; all \; \theta$ …............. (4.17)

and hence, $b^2(\theta) + \frac{[1+b'(\theta)]^2}{nI(\theta)} \leq \frac{1}{n} \;\; for \; all \; \theta$ …………… (4.18)

We shall then show that (4.18) ⇒ $b(\theta) \equiv 0$ for all θ. i.e δ is unbiased.

5. Since $|b(\theta)| \leq \frac{1}{\sqrt{n}}$ the function b is bounded.
6. From the fact that $1 + b'^2(\theta) + 2b'(\theta) \leq 1 \Rightarrow b'(\theta) \leq 0$ so that b is non-increasing.
7. Next, there exists a sequence of $\theta_i \to \infty$ and such that $b'(\theta_i) \to 0$

For suppose that $b'(\theta)$ were bounded away from 0 as $\theta \to \infty$, $say\ b'(\theta) \leq -\varepsilon\ for\ all\ \theta,$ then $b(\theta)$ can not be bounded

as $\theta \to \infty$, which contradicts 1.

8. Analogically it is seen that there exist a square $\theta_i \to -\infty$ and such that $b'(\theta_i) \to 0$ .Thus $b(\theta) \to 0$ as $\theta \to \pm\infty$ with inequality (4.18). Thus $b(\theta) \equiv 0$ follows from 2.
   Since $b(\theta) \equiv 0 \Rightarrow b'(\theta) = 0\ for\ all\ \theta \Rightarrow (4.16)\ as\ R(\theta,\delta) \leq \frac{1}{n} For\ all\ \theta\ and\ hence\ R(\theta,\delta) \equiv \frac{1}{n}$

This proves that $\bar{X}$ is admissible and minimax. This is unique admissible and minimax estimator. Because if δ' is any other estimator such that $R(\theta,\delta') \equiv \frac{1}{n}$. Then let $\delta^* = \frac{1}{2}(\delta + \delta')$

$$R(\theta,\delta^*) < \frac{1}{2}[R(\theta,\delta) + R(\theta,\delta')] = R(\theta,\delta)$$

Which contradicts that δ is admissible. Thus δ=δ' with prob. 1.

6.5 Equalizer Rules

The equalizer rule for exact minimax estimation and then proceeds to minimax hypothesis testing (also known as minimax detection).

The Equalizer Rule-

Suppose Θ is the parameter space and let d:Θ*Θ→ R^+be a specific loss function. The risk of an estimator $\hat{\theta}$ is defined as $E_\theta\left[d\left(\hat{\theta},\theta\right)\right]$, where the expection is taken over the iid random sample from the underlying distribution parameterized by the true parameter θ. Let πbe the prior distribution over the parameter space Θ. The Bayes risk of an estimator $\hat{\theta}$with respect to prior $\pi$is defined as-

$$R(\hat{\theta},\pi) = \int E_\theta \left[d(\hat{\theta},\theta)\right]d\pi(\theta)$$

The posterior risk of an estimator $\hat{\theta}$with respect to prior $\pi$is and data X is defined as-

$$r\left(\hat{\theta}/_X\right) = E_{\theta\sim\pi}\left[d(\hat{\theta},\theta)/X\right]$$

The Bayes rule estimator with respect to prior π is the estimator $\hat{\theta}$ that minimizes the posterior risk $r\left(\hat{\theta}/_X\right)$ at every X.

The **equalizer rule** asserts that an estimator is minimax if it is the Bayes rule with respect to some prior $\pi$and achieves the constant risk for all underlying parameterθ.

$\boldsymbol{Minimax\ strategy}$ $-$A minimax strategy for player 2 is a strategy δ^{M*}that minimizes the $sup_{\theta\epsilon\Theta}L(\theta,\delta^*)$ i.e. the strategy for which $sup_{\theta\epsilon\Theta}L(\theta,\delta^{M*}) = \inf sup_{\theta\epsilon\Theta}L(\theta,\delta^*)$
The R.H.S. is the minimax value of the game and denoted by $\bar{V}$.

$\boldsymbol{Maximin\ strategy}$- A maximin strategy for player 1 is a randomized strategy δ^M that maximizes$inf_{\theta\epsilon\Theta}L(\theta,a)$, i.e. the strategy for which

$$inf_{\theta\epsilon\Theta}L(\theta,\delta^M) = \sup inf_{\theta\epsilon\Theta}L(\theta,a)$$

The R.H.S. is the maximin value of the game and denoted by$\underline{V}$.

$\boldsymbol{Definition}$ $-$A strategy π_0 is equalizer for 1 if $L(\pi_0,a) = C\ (someconstant)\forall a \in A$. A strategy δ_0^* is an equalizer for player 2 if $L(\theta,\delta_0^*) = C'(someconstant)\forall\theta \in \Theta$.

Theorem- If both the player 1 and 2 have equalizer strategies, then the game has a value and the equalizer strategies are the maximin and the minimax strategies.

Proof- If π and $\delta^*$are the equalizer strategies then

$L(\theta,\delta^*) = K_1\forall\theta \in \Theta$, and $L(\pi,a) = K_2\forall a \in A$

$L(\pi, \delta^*) = E^{\pi} L(\theta, \delta^*) = E^{\pi} K_1 = K_1$

$L(\pi, \delta^*) = E^{\delta^*} L(\pi, a) = E^{\delta^*} K_2 = K_2$

Hence $K_1 = K_2$. Game has the value

Example: Binomial distribution.

Suppose $\boldsymbol{X \sim B(n, \theta)}$. Consider the Beta prior$\boldsymbol{\theta \sim Beta(\alpha, \beta)}$

The posterior distribution of θ conditioned on X is then $\boldsymbol{\theta / X \sim Beta(\alpha + x,\ \beta + n - x)}$

Under the squared error loss function $d(\widehat{\theta}, \theta)$=$(\hat{\theta} - \theta)^2$, the bayes rule is the posterior mean:

$$\hat{\theta}(\pi) = \frac{\alpha + x}{\alpha + \beta + n}$$

Taking $\rightarrow \beta \rightarrow \sqrt{n}/_2$, we have

$$R(\hat{\theta}(\pi), \theta) = \frac{1}{4(1 + \frac{1}{\sqrt{n}})^2}$$

which is a constant function with respect to the underlying parameter θ. Subsequently, by the equalizer rulewe claim that the minimax estimator for θ is

$\hat{\theta}$=$\frac{1}{(1+\sqrt{n})} + \frac{x}{(n+\sqrt{n})}$.

6.6 Multiple Decision Problems

A multiple decision problem is a problem in which only a finite set of actions (more than 2), is available.

(NOTE: For more details on this section please refer to Unit 1 of Block 1.)

6.7 Continuous Form of Bayes Theorem, Its Sequential Nature, Its Need in Decision Making

Consider a decision problem specified a parameter Θ whose value are in Θ (parameter space), a decision space D, and loss function L. we shall suppose that before the statistician chooses the decision in D, he will be permitted to observe sequentially the values of a sequence of r.v's $X_1, X_2, \dots$.... we shall suppose also that for any given value Θ=θ, these observations are independent and identically distributed. It is then said that the observations are a *sequential random sample*. We shall suppose that the conditional p.d.f. of each observation X_i when Θ=θ is $f(./\theta)$ and that the cost of observing the values X_i, in turn is C.

A sequential decision function or sequential decision procedure has two components. One component may be called as ***sampling plan*** or ***stopping rule***. The statistician first specifies whether a decision should choose without any observations or whether at least one observation should be taken. If at least one observation is to be taken, the statistician specifies, for every possible set of observed values $X_1 = x_1, X_2 = x_2, X_n = x_n (n \geq 1)$

whether sampling should stop and a decision in D chosen without further observations or whether another value X_{n+1} should be observed.

The second component of sequential decision procedure may be called a ***decision rule***. If no observations are to be taken, the statistician specifies a decision $d_0 \varepsilon D$ that is to be chosen. If at least one observation is to be taken, the statistician specifies the decision $d_n(x_1, \dots, x_n) \varepsilon D$ that is to be chosen for each possible set of observed values $X_1 = x_1, X_2 = x_2, X_n = x_n$ after which the sampling might be terminated.

Let S denote the sample space of any particular o bservation X_1. For n=1, 2… We shall let $S^n = SxSx \dots xS$ (with n factors) be the sample space of the n observations $X_1, X_2, \dots, X_n$ and we shall let S^∞ be the sample space of the infinite sequence of observations $X_1, X_2, \dots$

A sampling plan in which at least one observation is to be taken can be characterized by a sequence of subsets $B_n \varepsilon S^n$ (n=1, 2…) which have the following interpretations:

Sampling is terminated after the values $X_1 = x_1, \dots, X_n = x_n$

have been observed if $(x_1, \dots, x_n) \varepsilon B_n$. Another value x_{n+1} is observed if $(x_1, \dots, x_n) \notin B_n$. If there is some value r for which $B_r = S^r$ or more generally if $P[(x_1, \dots, x_n) \notin B_n]$for n=1, 2… r] =0 then the sampling must stop after at most r observations have been taken. The specification of the sets B_n for any value of n such that $n > r$ then become irrelevant never the less, it is convenient to assume that the sets B_n will be defined for all values of n.

Each stopping sets B_n can be regarded not only as a subset of S^n but also as the subset of S^r for any value of r> n and as a subset ofS^∞. When B_n is regarded as a subset of$S^r, r > n$, B_n is a cylinder set. In other words if $(x_1, \dots, x_n) \varepsilon B_n$ and if $(y_1, \dots, y_r)$ is any other set in S^r such that, $y_i = x_i$, i=1,2……..n then $(y_1, \dots, y_r) \varepsilon B_n$ regarded as of the values of the final r-n components.

Suppose that at least one observation is to be taken with a given sampling plan, and let N denote the random total number of observations which will be taken before sampling is terminated. We shall [N=n] denote the set of points $(x_1, \dots, x_n) \varepsilon S^n$ for which [N=n]. in other words, suppose that the value $X_1 = x_1, \dots, X_n = x_n$ are observed in sequence, then sampling will be terminated after the value $x_n$has been observed (and not before) if and only if $(x_1, \dots, x_n) \varepsilon [N = n]$. hence [N=1] = B_1 and for $n > 1$

$$[N = n] = (B_1 \cup B_2 \cup \dots \cup B_{n-1})^C \cap B_n$$

Similarly, we shall let $[N \leq n] = \bigcup_{i=1}^{n} [N = i]$ denote the subset of $S^n$for which $N \leq n$ the events $[N \leq n]$ and [N=n] involve only the observations $X_1, X_2, \dots, X_n$. Hence these events are subset of S^n. Also, they can be regarded as subsets of $S^r, r > n$. further more, events $[N > n] = [N \leq n]^C$involve the observations $X_1, X_2, \dots, X_n$, and it can be regarded as subsets of S^r for any value of r, $r \geq n$.

For any prior p.d.f ξ of θ, we shall let $f_n(./\xi)$ denote the marginal p.d.f of the observations $X_1, X_2, \dots, X_n$

$$f(x_1, \dots, x_n/\xi) = \int_\Theta f(x_1/\theta), \dots, f(x_n/\theta)\xi(\theta) d\nu(\theta) \dots\dots\dots\dots (6.1)$$

Furthermore, we shall let $f_n(./\xi)$ denote the marginal joint d.f of $X_1, X_2, \dots, X_n$. Hence, for any event$A \subset S_n$,

$$P[x_1, \dots, x_n \varepsilon A] = \int_A \; df_n\,(x_1, \dots, x_n/\xi) \;\; \dots\dots\dots\dots\dots \;\; (6.2)$$

We can write the following equation:

$$P[N \leq n] = \int_A \; dF_n\,(x_1, \dots, x_n/\xi) =$$

$$\int_{[N=1]} dF_1\,(x_1/\xi) + \int_{[N=2]} dF_2\,(x_1, x_2/\xi) + \int_{[N=3]} dF_3\,(x_1, x_2, x_3/\xi) + \cdots +$$
$$\int_{[N=n]} dF_n\,(x_1, x_2, \dots, x_n/\xi) \;\dots\dots\dots\dots\dots\; (6.3)$$

The decision rule of a sequential decision procedure is characterized by a decision rule $d_0 \varepsilon D$ and the sequence of functions $\delta_1, \delta_2, \dots$. with the following property: for any point$(x_1, \dots, x_n) \varepsilon S^n$, the function δ_n satisfies a decision, $\delta_n(x_1, \dots, x_n) \varepsilon D$. If the sampling plan specifies that an immediate decision in D is to be selected without any sampling then the decision $d_0 \varepsilon D$is chosen. If on the other hand, the sampling plan satisfies that at least one observation is to be taken and if the observed value $(x_1, \dots, x_n)$ satisfies the condition $(x_1, \dots, x_n) \varepsilon [N = n]$, then sampling is terminated and the decision , $\delta_n(x_1, \dots, x_n) \varepsilon D$ is chosen. The value of the function , δ_n need only be specified on the subset [N=n] $\subset S^n$. A procedure involving a fixed number of observations n can always be obtained by adopting a sampling plan in which [N=j] =Φ, the empty set for j=1… n-1 and in which [N=n] =S^n. In general we can also consider sampling plans for which the probability is 1 that sampling will eventually be terminated. In other words, we shall assume that,

$$P[N < \infty] = \lim_{n \to \infty} P[N \leq n] = 1 \;\dots\dots\dots\dots\dots\dots\; (6.4)$$

[It need not be assumed that there is some finite upper bound n such that$P[N \leq n] = 1$]

<u>Risk of a Sequential Decision Procedure</u>

The total risk $\rho(\xi, d)$ of a sequential decision procedure which at least one observation is to be taken is,

$$\rho(\xi, \delta) = E\{L[\theta, \delta_N(X_1, \dots, X_n)] + C_1 + C_2 + \cdots + C_N\}$$

$= \sum_{n=1}^{\infty} \int_{[N=n]} \int_{\Theta} L[\theta, \delta_n(X_1, \dots, X_n)]\, (\theta / x_1, \dots, x_n) d\nu(\theta) dF_n(x_1, \dots, x_n/\xi) + \sum_{n=1}^{\infty}(C_1 + C_2 + \cdots + C_N)\, P[N = n]$ …………………. (6.5)

Here $\xi(./ x_1, \dots, x_n)$is posterior p.d.f of Θ after the values $X_1 = x_1, \dots, X_n = x_n$ have been observed. Alternatively,

$$\rho(\xi, \delta) = \int_{\Omega} \left\{\int_{[N=n]} L[\theta, \delta_n(X_1, \dots, X_n)]\right\} \left[\prod_{i=1}^{n} f(x_i/\theta) d\mu(\mu)\right] \xi(\theta) d\nu(\theta) + \sum_{n=1}^{\infty} (C_1 + C_2 + \cdots + C_N)\, P[N = n] \dots \dots \dots \dots \dots \dots . (6.6)$$

In the development of theory of sequential statistical decision problem, we shall have little need to refer to any specified value ξ $(\theta / x_1, \dots, x_n)$ of the posterior p.d.f of Θ. However, we shall often have to refer to the entire posterior distribution as represented by its generalized p.d.f. therefore we shall denote the p.d.f simply by$\xi(x_1, \dots, x_n)$. $If\ \xi\ is\ prior\ distribution\ of\ \Theta.$ Where$X_1 = x_1, \dots, X_n = x_n$ is$\xi(x_1, \dots, x_n)$.

For every p.d.f of θ. Let $\rho_0(\Phi)$ be defined as follows:

$\rho_0(\Phi) = inf_{d\varepsilon D} \int_{\Omega} L[\theta, d]\Phi(\theta) d\nu(\theta)$ ……………… (6.7)

In other words, $\rho_0(\Phi)$ is the minimum risk from an immediate decision without any further observations when the p.d.f of θ is $\Phi(\theta)$.

A Bayes sequential decision procedure or an optimal sequential decision procedure is a procedure δ for which the risk ρ (ξ, δ) is minimized. Wherever a decision in D is chosen after sampling is terminated, that decision rule Bayes decision against the posterior distribution of Θ. For any such procedure δ which specifies that at least one observation is to be taken, we now have

$\rho(\xi, \delta) = E[P_0[\xi(x_1, \dots, x_n).] + C_1 + C_2 + \cdots + C_N]$ ………… (6.8)

Further, more for the procedure δ_0 which specifies that can immediate decision in D should be chosen without any observations we must have, $\rho(\xi, \delta_0) = \rho_0(\xi)$ ………………… (6.9)

Exp 6.1: $L(\theta_1, d_1) = L(\theta_2, d_2) = 0 \qquad \Theta = \{\theta_1, \theta_2\}, D = \{d_1, d_2\}$

$$L(\theta_1, d_2) = L(\theta_2, d_1) = b > 0$$

Suppose X is discrete r.v.'s for which

$$f_i(x) = P[X = x/\theta = \theta_i] \; i = 1,2$$

$$f_1(1) = 1 - \alpha, \qquad f_1(2) = 0, \qquad f_1(3) = \alpha \qquad 0 < \alpha < 1$$

$$f_2(1) = 0, \qquad f_2(2) = 1 - \alpha, \qquad f_2(3) = \alpha$$

Suppose the cost per observation is C, let the prior distribution of θ is $P[\theta = \theta_1] = \xi = 1 - P[\theta = \theta_2] \qquad \xi \leq \frac{1}{2}$

Solution: $\xi(\theta/x) = \frac{f(x/\theta)P[\theta=\theta]}{P[X=x]}$

$$\xi(\theta_1/1) = \frac{(1-\alpha)\xi}{(1-\alpha)\xi + 0} = 1 \qquad \xi(\theta_1/1) = 0$$

$$\xi(\theta_1/3) = \frac{f(3/\theta_1)P[\theta = \theta_1]}{f(3/\theta_1)P[\theta = \theta_1] + f(3/\theta_2)P[\theta = \theta_2]}$$

$$= \frac{\alpha\xi}{\alpha\xi + \alpha(1-\xi)} = \xi$$

Similarly, $\xi(\theta_2/1) = 0, \quad \xi(\theta_2/2) = 1, \quad \xi(\theta_2/3) = (1 - \xi)$

Thus, after an observation has been taken, either the value of θ becomes known or else the distribution of θ remains good as it was before the observation was taken.

$$\rho_0(\xi) = inf_d\{L(\theta_1, d_1)\xi + L(\theta_2, d_1)(1-\xi), L(\theta_1, d_2)\xi + L(\theta_2, d_2)(1-\xi)\}$$

$= inf_d\{b(1-\xi), b\xi\}$ Without any observation is taken.

$$= b\xi \qquad since, \xi \leq \frac{1}{2}$$

If the Bayes decision is chosen when $P[\theta = \theta_1] = \xi$, the expected loss is bξ.

If one observation is taken then the expected loss will be

$$E\ \rho_0(\xi(X)), \text{where } \xi(X) = P[\theta = \theta/X = x]$$

$\rho_0(1) = \rho_0(\xi(1))$

$= inf_d\{L(\theta_1, \delta(1))P[\theta = \theta_1/X = 1] + L(\theta_2, \delta(1))P[\theta = \theta_2/X = 1]\}$

$$= inf_d\{0, \mathrm{b}\} = 0$$

Now, $L(\theta_1, \delta(1))P[\theta = \theta_1/X = 1] + L(\theta_2, \delta(1))P[\theta = \theta_2/X = 1]$

$$\begin{aligned} &= 0 && if\,\delta(1) = d_1 \\ &= b && if\,\delta(1) = d_1 \end{aligned}$$

Similarly, $\rho_0(2) = 0 \quad and \quad \rho_0(3) = b\xi$

$$E\rho_0(X) = 0P[X = 1] + 0P[X = 2] + \mathrm{b}\xi P[X = 3] = \mathrm{b}\xi\alpha$$

The expected loss $E\rho_0(X_1, \dots, X_n) = \mathrm{b}\xi\alpha^{\mathrm{n}}$ when the Bayes decision is chosen after n observations $X_1, \dots, X_n$ have been taken,

$\rho_n = \mathrm{b}\xi\alpha^{\mathrm{n}} + \mathrm{Cn}$ Total risk for the optimal procedure when exactly n observations taken, assume ρ (1) <ρ (0)

$\frac{d}{dn}\rho(\mathrm{n}) = 0 \Rightarrow n^* = [\log\frac{\mathrm{b}\xi\log(\frac{1}{\alpha})}{\mathrm{C}}]\frac{1}{\log(\frac{1}{\alpha})}$ .... (6.10)

and $\rho(\mathrm{n}^*) = \frac{\mathrm{C}}{\log(\frac{1}{\alpha})}\left[1 + \log\frac{\mathrm{b}\xi\log(\frac{1}{\alpha})}{\mathrm{C}}\right]$ .... (6.11)

Wolfowitz Generalization of FCR bound and Sequential estimation and Testing:

A sequential provides a set of stopping rules $\{R_n(X_1, \dots, X_n); n = 1,2 \dots \dots\}$ which are $\mathfrak{B}^{(n)}$ designate the Borel σ-field on $\mathfrak{x}^{(n)}$, n-dimensional Euclidian space; assigning to $(X_1, \dots, X_n)$ an integral value so that if $R_n(X_1, \dots, X_n) = n$, we terminate sampling after the n^{th} observation otherwise, $X_{n+1}$is observed. Consider the σ-field$\mathfrak{B}_1 \subset \mathfrak{B}_2 \subset \cdots$ generated by $X_1, .., (X_1, \dots, X_n)$ a stopping rule R for a sequential procedure can be conveniently described by a sequence of sets $\{R_n; n = 1,2, \dots.\}$ where, $R_n \varepsilon \mathfrak{B}_n$ for each n=1,2,….. Sampling is continued as by as consecutive vectors $(X_1, \dots, X_n)$, n=1,2,….. do not enter one of the sets R_n. In another words, the sample size N (a random variable) is N= least integral n, n≥ 1 such that $(X_1, \dots, X_n)\varepsilon R_n$

Define sets,$\overline{R_n} = \begin{array}{ll} R_1 & if n = 1 \\ \overline{R_1} \cap \overline{R_2} \cap \dots \cap R_n & if n \geq 2 \end{array}$

The sets $\overline{R_n}$ is the set of all sample points which leads to stopping at N=n. The estimation rule for estimating a function g $(P_1, P_1, \dots)$ is given by a srquence of functions $\widehat{g_1}, \widehat{g_2}, \dots$ such that $\widehat{g_n} \varepsilon \mathfrak{B}_n$ for all n=1,2… and if N=n then the estimate of g is $\widehat{g_n}$.

<u>Lemma 9.1</u>: $[\boldsymbol{wald's\ equation}]$: let$(X_1, \dots, X_n \dots)$ be a sequence of i.i.d random variables, distributed with some distribution, satisfying E$|X| < \infty$. For any sequential rule yielding EN$< \infty$

$$E(\textstyle\sum_{i=1}^{N} X_i) = E(X)EN \text{ …………………. (9.2)}$$

<u>Proof</u>: let $(R_1, R_2, \dots)$ be the sequence of stopping regions. Then,

$$E(\textstyle\sum_{i=1}^{N} X_i) = \sum_{n=1}^{\infty} \int_{\overline{R_n}} \sum_{i=1}^{n} x_i \left(\prod_{i=1}^{n} dF(x_i)\right) \text{ ………… (9.2)}$$

Now, $EX_i = \sum_{n=1}^{\infty} \int_{\overline{R_n}} (x_i) \prod_{i=1}^{n} dFx_i$

$$= \sum_{n=1}^{i-1} \int_{\overline{R_n}} x_i \prod_{i=1}^{n} dF(x_i) + \sum_{n=i}^{\infty} \int_{\overline{R_n}} x_i \prod_{i=1}^{n} dF(x_i)$$

$$= E\{X_i I[N < i]\} + E\{X_i I[N \geq i]\}$$

$$\sum_{n=i}^{\infty} \int_{\overline{R_n}} x_i \prod_{i=1}^{n} dF(x_i) = E\{X_i I[N \geq i]\} = P[N \geq i] E[X_i / N \geq i]$$

Since $[N \geq i]$ is $\mathfrak{B}_{i-1}$ measure and $\mathfrak{B}_0 = \mathfrak{B}$, therefore X_i is independent of $[N \geq i]$. thus

$$E[X_i / N \geq i] = E(X_i)$$

$$\sum_{n=i}^{\infty} \int_{\overline{R_n}} x_i \prod_{i=1}^{n} dF(x_i) = P[N \geq i] E(X_i)$$

$$= P[N \geq i] E(X) \text{ .... (9.3)}$$

Now from (9.1)

$\sum_{n=i}^{\infty} \int_{\overline{R_n}} \sum_{i=1}^{n} x_i \prod_{i=1}^{n} dF(x_i) = \sum_{i=1}^{\infty} \sum_{n=i}^{\infty} \int_{\overline{R_n}} x_i \prod_{i=1}^{n} dF(x_i)$ .... (9.4)

(This is permitted as E$|X| < \infty$)

$= \sum_{i=1}^{\infty} P[N \geq i] E(X)$ From (9.3)

$$= EX \sum_{i=1}^{\infty} P[N \geq i] = E(X) EN$$

$E(\sum_{i=1}^{n} X_i) = E(X) EN\#$

Alternative Proof: Define a r.v. Y_i such that

$Y_i = 1, if\ no\ decision\ is\ reached\ up\ to\ (i-1)th\ stage, i.e.\ if\ N > (i-1)$

$0 \quad otherwise.$

Clearly, Y_i depends only on $X_{1,} X_2, \ldots\ldots, X_{i-1}$ $\ and\ does\ not\ depend\ on\ X_i$. Also

$$S_N = \sum_{n=1}^{\infty} X_n Y_n$$

Hence $E(S_N) = E(\sum_{n=1}^{\infty} X_n Y_n)$ (9.5)

Now,

$$\sum_{n=1}^{\infty} E|X_nY_n| = \sum_{n=1}^{\infty} E|X_n|E|Y_n| \; (because\ X_n\ and\ \ Y_n\ are\ independent)$$

$$= E|X_1|\sum_{n=1}^{\infty} E|Y_n| = E|X_1|\sum_{n=1}^{\infty} P[N \geq n] \; (becasuseE\,|Y_n| = P[Y_n = 1] = P[N \geq n])$$

$$= E|X_1|\sum_{n=1}^{\infty}\sum_{k=n}^{\infty} P[N = k] = E|X_1|\sum_{n=1}^{\infty} nP[N = n]$$

$$= E|X_1||E(N)| < \infty$$

Therefore, $E(S_N)$exists and we may change the order of operation of expectation and summation sign in (9.5). Hence,

$$E(S_N) = E\left(\sum_{n=1}^{\infty} X_nY_n\right) = \sum_{n=1}^{\infty} E(X_nY_n) = E(X_1)\sum_{n=1}^{\infty} E(Y_n)$$

$$= E(X_1)\sum_{n=1}^{\infty} P[N \geq n] = E(X_1)E(N)$$

Note: Lemma 9.1 holds if only we assume $E(X_n) = \mu\ and\ E(N) < \infty$ and the assumption that $X_i'sare\ i.i.d.is\ not\ necessary.$

Lemma 9.2: Let$(X_1, \ldots, X_n)$ be a sequence of i.i.d random variables, having a common d.f. F(x) with mean zero and variance

$\sigma^2, 0 < \sigma^2 < \infty$for any sequential stopping rule with E(N)$< \infty$, if

$E\{(\sum_{i=1}^{N}|X_i|)^2\} < \infty\ \ then, E\{(\sum_{i=1}^{N} X_i)^2\} = \sigma^2 EN$ ………… (9.5)

Proof: As before,

$$E\left\{\left(\sum_{i=1}^{N} X_i\right)^2\right\} = \sum_{n=1}^{\infty} \int_{\overline{R_n}} \left(\sum_{i=1}^{n} x_i\right)^2 \prod_{i=1}^{n} dF(x_i)$$

$$=\sum_{n=1}^{\infty}\int_{\overline{R_n}}\left\{\sum_{i=1}^{n}{x_i}^2+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x_ix_j\right\}\prod_{i=1}^{n}dF(x_i)$$

$$=\sum_{n=1}^{\infty}\int_{\overline{R_n}}\left(\sum_{i=1}^{n}{x_i}^2\right)\prod_{i=1}^{n}dF(x_i)+2\sum_{n=1}^{\infty}\int_{\overline{R_n}}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x_ix_j\right)\prod_{i=1}^{n}dF(x_i)$$

$= \sigma^2 EN + 2\sum_{n=1}^{\infty}\int_{\overline{R_n}}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x_ix_j\right)\prod_{i=1}^{n}dF(x_i)$ By Lemma 9.1

Now

$$\sum_{n=1}^{\infty}\int_{\overline{R_n}}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x_ix_j\right)\prod_{i=1}^{n}dF(x_i)=\sum_{i=2}^{\infty}\sum_{j=1}^{i-1}\sum_{n=i}^{\infty}\int_{\overline{R_n}}x_ix_j\prod_{i=1}^{n}dF(x_i)$$

But

$\sum_{n=i}^{\infty}\int_{\overline{R_n}}x_ix_j\prod_{i=1}^{n}dF(x_i)=P[N\geq i]E[X(x)/N\geq i]$ for j<i, (i=1, 2, 3…) as X_i is independent $[N\geq i]$

$= P[N\geq i]EX_iE\left[X_j/N\geq i\right]=0\quad for$ j < i, (i = 1, 2, 3 ...)

The rearrangement is guaranteed by condition $E\{(\sum_{i=1}^{N}|X_i|)^2\}<\infty$

Then $E\{(\sum_{i=1}^{N}X_i)^2\}=\sigma^2EN\#$

Alternative Proof: Let Y_i be defined as in Alternative proof of Lemma 9.1. Then

$$E(S_N)^2=E\left\{\left(\sum_{i=1}^{\infty}X_iY_i\right)\right\}\left\{\left(\sum_{j=1}^{\infty}X_jY_j\right)\right\}$$

$$= E\left(\sum_{i=1}^{\infty}X_i^2Y_i^2+\sum_{i\neq j}\sum_{j}X_iY_iX_jY_j\right)\qquad(9.6)$$

$$E|S_N^2| = E\left(\sum_{i=1}^{\infty} X_i^2 Y_i^2 + \sum_{i\neq j}\sum_{j} |X_i X_j| Y_i Y_j\right)$$

$$= E\left(\sum_{i=1}^{N} |X_i|\right)^2 < \infty \ (by\ assumption).$$

Hence the order of operation of summation and expectation in (9.6) can be interchanged. Now

$$E\left(\sum_{i=1}^{\infty} X_i^2 Y_i^2\right) = E(X_1^2)E\left(\sum_{i=1}^{\infty} Y_i^2\right) = \sigma^2 E\left(\sum_{i=1}^{\infty} Y_i\right) = \sigma^2 E(N)\ (by\ Lemma 9.1)$$

Again

$$E\left(\sum_{i\neq j}\sum_{j} X_i Y_i X_j Y_j\right) = 2E\left(\sum_{i>j}^{\infty}\sum_{j}^{i-1} X_i Y_i X_j Y_j\right) = 2\sum_{i=2}^{\infty}\sum_{j=1}^{i-1} E\left(X_i X_j Y_i\right)$$

$$= =2\sum_{i=2}^{\infty}\sum_{j=1}^{i-1} E\left\{Y_i E\{X_i X_j / Y_i\}\right\} = 2\sum_{i=2}^{\infty}\sum_{j=1}^{i-1} E\{Y_i E(X_i) E\left(X_j / Y_i\right)\} =$$

0

as X_j *and* Y_i *are independent of* X_i. #

Generalization of FCR bound for Sequential estimation

Theorem 9.1: [$\boldsymbol{wolfowitz}$]: Let$(X_1, \dots, X_n, \dots.)$ be a sequence of i.i.d random variables, whose common density $f(x;\ \theta)$ with respect to measure μ belong to a family $\psi = \{f(.;\ \theta): \theta \varepsilon \Theta\}$ on which the following regularity conditions are satisfied:

1. Θ contains an interval in a Euclidian k-space.
2. $f(x;\ \theta)$is differentiable w.r.to θ on Θ.
3. $\int\left|\frac{\partial}{\partial\theta} f(x;\ \theta)\right| d\mu < \infty\ for\ all\ \theta\varepsilon\Theta.$
4. $0 < \int\left[\frac{\partial}{\partial\theta} log f(x;\ \theta)\right]^2 f(x;\ \theta) d\mu < \infty\ for\ all\ \theta\varepsilon\Theta.$
5. $For\ each\ n = 1,2, \dots\dots..and\ all\ \theta$

$$\int [\sum_{i=1}^{n} \frac{\left|\frac{\partial}{\partial\theta} f(x_i;\,\theta)\right|}{f(x_i;\,\theta)}]^2 \prod_{i=1}^{n} dF(x_i) < \infty$$

$$or \int [\sum_{i=1}^{n} \left|\frac{\partial}{\partial\theta} log f(x_i;\,\theta)\right|]^2 \prod_{i=1}^{n} dF(x_i) < \infty$$

Let $(R_n, n = 1,2 \ldots)$ be the sequence of stopping regions associated with a given sequential procedure. Let $g(\theta)$ be an estimable and differential function on Θ. Let $\hat{g}(X_1, \ldots, X_n, \ldots)$ be unbiased estimator of $g(\theta)$ satisfying the following conditions:

6. $\int |\hat{g}(x_1, \ldots, x_n)| \frac{\partial}{\partial\theta} \prod_{\nu=1}^{n} f(x_\nu;\,\theta) \prod_{\nu=1}^{n} d\mu(x_\nu) < \infty \; for\ each$
$n = 1,2 \ldots$
7. $\sum_{n=1}^{\infty} \frac{d}{d\theta} g_n(\theta)$ converges uniformly on Θ, where

$$g_n(\theta) = \int_{\overline{R_n}} \hat{g}(x_1, \ldots, x_n) \prod_{\nu=1}^{n} dF(x_\nu)$$

$then\ Var_\theta\{\hat{g}(X_1, \ldots, X_n, \ldots)\} \geq \frac{[g'(\theta)]^2}{I(\theta)E(N)}$ .... (9.6)

$for\ all\ \theta, provided\ EN < \infty$

<u>Proof</u>: Let N be the sample size associated with the given sequential procedure. Let $S(X_i;\,\theta) = \frac{d}{d\theta} log f(X_i;\,\theta); i = 1,2 \ldots$

These are i.i.dr.v's and 1-4 guarantee that E $S(X_i;\,\theta) = 0$ and $I(\theta) = E[S^2(X_i;\,\theta)] < \infty$ by condition 4 and the assumption $E(N) < \infty \Rightarrow by\ Lemma\ 9.1$

$E[\sum_{i=1}^{N} S(X_i;\,\theta)] = E(N) ES(X_i;\,\theta) = 0 \;\; for\ all\ \theta$ .... (9.7)

Furthermore, according to condition 5

$E[\sum_{i=1}^{N} |S(X_i;\,\theta)|]^2 < \infty$ .... (9.8)

$E[\{\sum_{i=1}^{N} S(X_i;\,\theta)\}^2] = E(N) ES^2(X, \theta) = E(N) I(\theta)$ .... (9.8)

Consider the expectation,

$$\mathrm{E}\left\{\hat{g}(X_1, \dots, X_n \dots) \sum_{i=1}^{N} \mathrm{S}(X_i;\, \theta)\right\} \qquad \theta \varepsilon \Theta$$

Where $\hat{g}(X_1, \dots)$ is unbiased estimator of $g(\theta)$. According to (9.7) and by Schwartz inequality we have

$$\mathrm{E}\left\{\hat{g}(X_1, \dots, X_N) \sum_{i=1}^{N} \mathrm{S}(X_i;\, \theta)\right\} \leq \left[\mathrm{E}\left\{\left(\hat{g}(X_1, \dots, X_N) - g(\theta)\right)^2\right\} \mathrm{E}\left\{\left(\sum_{i=1}^{N} \mathrm{S}(X_i;\, \theta)\right)^2\right\}\right]^{\frac{1}{2}}$$

For all θ ε Θ ……………… (9.10)

The quantity $\mathrm{E}\left(\hat{g}(X_1, \dots, X_{N\dots}) - g(\theta)\right)^2$ is the variance of $\hat{g}(X_1, \dots, X_n \dots.)$ under the sequential procedure. Further 6 & 7 allow the differentiation under the integral sign in,

$$g'(\theta) = \frac{d}{d\theta} \sum_{n=1}^{\infty} \int_{\widehat{R_n}} \hat{g}(x_1, \dots, x_n) \prod_{\nu=1}^{n} dF(x_\nu)$$

$$= \sum_{n=1}^{\infty} \int_{\widehat{R_n}} \hat{g}(x_1, \dots, x_n) \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i;\, \theta) \prod_{i=1}^{n} d\mu(x_i)$$

$$= \sum_{n=1}^{\infty} \int_{\widehat{R_n}} \hat{g}(x_1, \dots, x_n) \left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} log f(x_i;\, \theta)\right) \prod_{i=1}^{n} f(x_i;\, \theta)\, d\mu(x_i)$$

$$= \sum_{n=1}^{\infty} \int_{\widehat{R_n}} \hat{g}(x_1, \dots, x_n) \left(\sum_{i=1}^{n} S(x_i;\, \theta)\right) \prod_{i=1}^{n} dF(x_i)$$

$= E[\hat{g}(X_1, \dots, X_n \dots) \sum_{i=1}^{N} S(X_i;\, \theta)]$ ……………………… (9.11)

From (9.9) (9.10) & (9.11)

$$Var_\theta \hat{g}(X_1, \dots \dots) \geq \frac{E^2[\hat{g}(X_1, \dots, X_{n\dots}) \sum_{i=1}^{Nn} S(X_i;\, \theta)]}{I(\theta) E(N)}$$

$$= \frac{[g'(\theta)]^2}{I(\theta) E(N)} \#$$

Optimality Criterion of Sequential Procedure

1. Subject to the condition $E_\theta(N) \leq m$ (m is a fixed integral bound) for all θ, minimize the variance of the best unbiased estimator that is, $E_\theta(\widehat{g_N} - g)^2$ uniformly in θ (if such an estimator exist.)
2. Subject to the condition $E(\widehat{g_N} - g)^2 \leq v < \infty (fixed\ finite\ positive\ value) for\ all\ \theta,$ minimize expected sample size $E_\theta(N)$.
3. Minimizes the expected cost of sampling plus expected loss, that is, $CE_\theta(N) + E_\theta(\widehat{g_N} - g)^2$

Generally, there is no sequential estimator that can satisfy 3 uniformly in θ. In case 2, De$Groot(1959)\ and\ Wasan(1964)$ have shown that a fixed sample size procedure in the binomial case does not minimize $E_\theta(N)$ w.r.to all sequential procedure uniformly in θ, 0$< \theta < 1$ subject to the condition that $sup_{0<\theta<1} var_\theta(\hat{g}) \leq \frac{1}{4m}$.

Sequential Estimation of the Mean of Normal Population

Let$(X_1, \dots, X_n)$ be i.i.d r.v's with mean μ and varianceσ^2, both unknown as an estimate of μ, we choose$\bar{X}_n$, the sample mean. The problem now is to choose n. Let us assume that the loss incurred is A$|\bar{X}_n - \mu|$, where A>0, is known constant and let each observation cost one unit. Then we wish to choose n to minimize,

$$EL(n) = E\{A|\bar{X}_n - \mu| + n\} \text{ ……………………. (9.12)}$$

We have, $E\sqrt{n}\frac{|\bar{X}_n - \mu|}{\sigma} = \sqrt{\frac{2}{\pi}}$

So that $EL(n) = AE\left(\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}\right)\frac{\sigma}{\sqrt{n}} + n$

$$= A\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}} + n \text{ ……………….. (9.13)}$$

Treating as continuous function n we have for minimax,

$$-A\sqrt{\frac{2}{\pi}}\frac{\sigma}{2(n)^{\frac{3}{2}}}+1=0 \Rightarrow n_0=\left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}} \text{…………………. (9.14)}$$

At the value n that minimizes (9.13), for this value of n

$$\nu(\sigma)=EL(n_0)=A\sqrt{\frac{2}{\pi}}\sigma\left(\frac{\sqrt{2\pi}}{A\sigma}\right)^{\frac{1}{3}}+\left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}}$$

$$=A\sqrt{\frac{2}{\pi}}\sigma\left(\frac{\sqrt{2\pi}}{A\sigma}\right)^{\frac{-1}{3}}+\left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}}$$

$$=\frac{A\sqrt{\frac{2}{\pi}}\sigma+\frac{A\sigma}{\sqrt{2\pi}}}{\left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}}}=3\left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}}=3n_0 \text{ ……….. (9.15)}$$

So that the loss due to the error of estimation is thrice the size of the sample, that is thrice the cost of sampling. Of course, this presupposes the knowledge of σ. If we do not know σ, we cannot compute n_0.

When σ is not known, we have the following sequential sampling procedure R:

$$N=\ least\ n\ , n\geq 2\ where\quad n\geq\left(\frac{As_n}{\sqrt{2\pi}}\right)^{\frac{2}{3}} \text{ ……………. (9.16)}$$

Where, ${s_n}^2=\frac{\sum(x_i-\overline{x_n})^2}{n-1}$, $\overline{x_n}=\frac{1}{n}\sum_{i=1}^{n}x_i$

We may write this inequality,

$$N=\ first\ n\ , n\geq 2\ when\ \sum_{i=1}^{n}(x_i-\overline{x_n})^2\leq\frac{2\pi}{A^2}(n-1)n^3 \text{ …… (9.17)}$$

<u>Lemma 9.3</u>: Rule R terminates with probability 1.

<u>Proof</u>: It is sufficient to show that,

$$\left(\frac{As_n}{\sqrt{2\pi}}\right)^{\frac{2}{3}}\xrightarrow{P} n_0 \quad i.e\ \lim_{n\to\infty}P\left[\left|\left(\frac{As_n}{\sqrt{2\pi}}\right)^{\frac{2}{3}}-n_0\right|\leq\varepsilon\right]=1$$

Or $\lim_{n\to\infty} P[\left|\left(\frac{As_n}{\sqrt{2\pi}}\right)^{\frac{2}{3}} - n_0\right| > \varepsilon] = 0$

Now $\quad \lim_{n\to\infty} P\left[\left|\left(\frac{As_n}{\sqrt{2\pi}}\right)^{\frac{2}{3}} - \left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}}\right| > \varepsilon\right]$

$$= \lim_{n\to\infty} P\left[\left|\left(\frac{s_n}{\sigma^2}\right)^{\frac{1}{3}} - 1\right| > \left(\frac{\sqrt{2\pi}}{A}\right)^{\frac{2}{3}} \varepsilon\right] \quad \ldots\ldots\ldots\ldots\ldots(9.18)$$

Since, $\lim_{n\to\infty} P\left[\left|\frac{s_n{}^2}{\sigma^2} - 1\right| > \left(\frac{\sqrt{2\pi}}{A}\right)^{\frac{2}{3}} \varepsilon\right] \leq \lim_{n\to\infty} \frac{2}{(n-1)} \varepsilon^2 \left(\frac{A\sigma}{\sqrt{2\pi}}\right)^{\frac{2}{3}} = 0$

As $\quad \frac{s_n{}^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$ therefore (9.18) tends to zero as n→ ∞.

<u>Lemma 9.4</u>: For any fixed n, $\overline{X_n}$ is independent of $S_2^2, S_3^2, \ldots, S_n^2$ and hence,

$$P\left[\sqrt{n}\left(\frac{\overline{X_n}-\mu}{\sigma}\right) \leq t/s_2^2, \ldots, s_n^2\right] = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad \ldots\ldots\ldots (9.19)$$

<u>Proof</u>: Define $U_i = \frac{X_i-\mu}{\sigma}$ $\; i = 1,2, \ldots, n$

Then $U_i \sim N(0,1)$ r.v's and independent i=1,2…..

Let us write,

$$y_i = \frac{u_1 + u_2 + \cdots + u_i - iu_{i+1}}{\sqrt{\mathrm{i}(\mathrm{i}+1)}}, \mathrm{i} = 1,2, \ldots, \mathrm{n}-1$$

$$y_n = \sqrt{\mathrm{n}}\bar{\mathrm{u}} \quad \text{where } \bar{\mathrm{u}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} u_i$$

$$cov(Y_i, Y_j) = E\left[\frac{U_1 + U_2 + \cdots + U_i - iU_{i+1}}{\sqrt{\mathrm{i}(\mathrm{i}+1)}} . \frac{U_1 + U_2 + \cdots + U_j - jU_{j+1}}{\sqrt{\mathrm{j}(\mathrm{j}+1)}}\right]$$

$$= E\left[\frac{(\mathrm{U}_1^2 + \mathrm{U}_2^2 + \cdots + \mathrm{U}_\mathrm{i}^2) - \mathrm{i}\mathrm{E}\mathrm{U}_{\mathrm{i}+1}^2}{\sqrt{\mathrm{i}(\mathrm{i}+1)(\mathrm{j}+1)}}\right] = \mathrm{i} - \mathrm{i} = 0$$

$$EY_i = 0\ , var(Y_i) = \frac{E\mathrm{U}_1^2 - \mathrm{i}^2\mathrm{EU}_{\mathrm{i+1}}^2}{\mathrm{i(i+1)}} = \frac{i+\mathrm{i}^2}{i(i+1)} = 1$$

$Y_i\ are\ i.i.d\ N(0,1)\ i = 1,2,\dots,n$

$$S_i^2 = \frac{1}{i-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

$$= \frac{\sigma^2}{i-1}\sum_{j=1}^{i-1}{Y_j}^2 = \frac{\sigma^2}{i-1}\left({Y_1}^2 + \cdots + {Y_{i-1}}^2\right), i = 2,3,\dots,n$$

It follows that Y_n is independent of S_i^2 for i=2,…,n this is the same as saying $\overline{X_n}$ is independent of $S_2^2, S_3^2, \dots, S_n^2$.

Let us now compute the average loss for R.

$$L(N) = A\sqrt{N}\left|\frac{\overline{X_n} - \mu}{\sigma}\right|\frac{\sigma}{\sqrt{N}} + N$$

$$EL(N) = \sum_{n=2}^{\infty} P[N = n]E[L(N)/N = n]$$

$$= \sum_{n=2}^{\infty} P[N = n]E[A\sqrt{N}\left|\frac{\overline{X_n} - \mu}{\sigma}\right|\frac{\sigma}{\sqrt{N}} + N/N = n]$$

$$= \sum_{n=2}^{\infty} P[N = n]AE\left[\sqrt{N}\left|\frac{\overline{X_n} - \mu}{\sigma}\right|\frac{\sigma}{\sqrt{N}} + N/N = n\right] + E(N)$$

$$= \sum_{n=2}^{\infty} P[N = n](A\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{N}}) + E(N)$$

$$= A\sqrt{\frac{2}{\pi}}\sigma E\left(N^{\frac{-1}{2}}\right) + E(N)$$

$$= 2n_0^{\frac{3}{2}}E\left(N^{\frac{-1}{2}}\right) + E(N)$$

Proposition: For large $n_0 P[N \leq n] \geq \frac{1}{2}$

Proof: We have, $P[N \leq n] \geq P\left[{Y_1}^2 + {Y_2}^2 + \cdots + {Y_{n-1}}^2 \leq \frac{(n-1)n^3}{n_0^3}\right]$ for $n = n_0$

$$P[N \leq n] \geq P\left[{Y_1}^2 + {Y_2}^2 + \cdots + {Y_{n_0-1}}^2 \leq n_0 - 1\right]$$

$$= P\left[\chi^2_{(n_0-1)} \leq n_0 - 1\right]$$

$$= P\left[\chi^2_{(n_0-1)} - \overline{n_0 - 1} \leq 0\right]$$

$= P[Z \leq 0] = \frac{1}{2}$ Where, $Z \sim N(0,1)\#$

Theorem 9.2: Let$(Z_1, \dots, Z_n)$ be i.i.d- r.v's such that $P[Z_j = 0] \neq 1$ set

$S_n = Z_1 + Z_2 \dots + Z_n$ and for two constants C_1, C_2 *with* $C_1 < C_2$, define the random quantity N as the smallest n for which $S_n \leq C_1$ *or* $S_n \geq C_2$, *set* $N = \infty$ *if* $C_1 < S_n < C_2$ *for all* n. thus there exist $C > 0$ *and* $0 < \rho < 1$ such that,

$P[N > n] \leq C\rho^n$ *for all* n. ………………. (9.20)

Proof: The assumption $P[Z_j = 0] \neq 1$ implies that $P[Z_j > 0] > 0$. Let us suppose that $P[Z_j > 0] > 0$ then there exists $\varepsilon > 0$ such that $P\left[Z_j > \varepsilon\right] = \delta > 0$ in fact if $P\left[Z_j > \varepsilon\right] = 0$ *for* $\forall > \varepsilon$, then in particular $P\left[Z_j > \frac{1}{n}\right] = 0$ *for all* n. but $P\left[Z_j > \frac{1}{n}\right] \uparrow P[Z_j > 0]$ and we have $0 = \lim_n P\left[Z_j > \frac{1}{n}\right] = P[Z > 0]$ which is a contradiction.

Thus for $P\left[Z_j > 0\right] > 0$ *we have* $P\left[Z_j > \varepsilon\right] = \delta > 0$ ……… (9.21)

With C_1, C_2 and ε, there exist a positive integer m such that,

$m\varepsilon > C_2 - C_1$ ……………….. (9.22)

For such m we have,

$\bigcap_{j=k+1}^{k+m}[Z_j > \varepsilon] \subseteq [\sum_{j=k+1}^{k+m} Z_j > m\varepsilon] \subseteq [\sum_{j=k+1}^{k+m} Z_j > C_2 - C_1]$ …. (9.23)

$$P[\sum_{j=k+1}^{k+m} Z_j > C_2 - C_1] \geq P\{\bigcap_{j=k+1}^{k+m}[Z_j > \varepsilon]\}$$

$$= \prod_{j=k+1}^{k+m} P[Z_j > \varepsilon] = \delta^m, as\ Z_j's\ are\ independent\ .$$

Clearly,

$$S_{km} = \sum_{j=0}^{k-1}[Z_{jm+1} + \cdots + Z_{(j+1)m}]$$

Now we assert that, $C_1 < S_i < C_2, i = 1,2, \dots, km \Rightarrow$

$Z_{jm+1} + \cdots + Z_{(j+1)m} \leq C_2 - C_1, j = 1,2, \dots, k-1$ ………. (9.24)

This is because, if for some $j = 1,2, \dots, k-1$ we suppose that $Z_{jm+1} + \cdots + Z_{(j+1)m} > C_2 - C_1$, this inequality together

$S_{jm} > C_1$ would imply $S_{(j+1)m} > C_2$, which is a contradiction to the first part of (9.24).

$$[N \geq km+1] \subseteq [C_1 < S_j < C_2, j = 1,2, \dots, km]$$

$$\subseteq [Z_{jm+1} + \cdots + Z_{(j+1)m} \leq C_2 - C_1]$$

$$P[N \geq km+1] \leq \prod_{j=0}^{k-1}[Z_{jm+1} + \cdots + Z_{(j+1)m} \leq C_2 - C_1]$$

$$\leq (1-\delta^m)^k$$

Thus, $P[N \geq km+1] \leq (1-\delta^m)^k = \frac{\left[(1-\delta^m)^{\frac{1}{m}}\right]^{mk+1}}{1-\delta^m} = C\rho^{mk+1}$

$$Put\ \ C = \frac{1}{1-\delta^m} \quad ,\rho = (1-\delta^m)^{\frac{1}{m}} \ \ ,0 < \rho < 1, C > 0$$

$thus, \qquad P[N \geq n] \leq C\rho^n$ #

Theorem 9.3: Let $M_\theta(t) = M_\theta(e^{tz})$ be the m.g.f of Z, and let it be assumed to exist for all t, where $Z = log\frac{f(x,\theta_1)}{f(x,\theta_0)}$ then a necessary and sufficient condition that there exist a $(t = t_0 \neq 0)$ such that $M_\theta(t_0) = 1$ is that $E_\theta(Z) \neq 0$ and that Z takes on both positive and negative values with positive probability.

Proof: To prove the sufficiency, we observe that

$M_\theta{}''(t) = E_\theta(Z^2 e^{tZ}) > 0$ Unless Z=0 [since$M_\theta(t)$ exists for all t, it is differentiable any number of times]. Thus $M_\theta(t)$is convex function of t. Now by assumption there exists a value Z’$>$ 0 such that $P_\theta[Z > Z'] = u > 0$, therefore $t > 0$ implies

$$M_\theta(t) = E_\theta(e^{tZ}) > e^{tZ'}P_\theta[Z > Z'] = ue^{tZ'} \quad \ldots\ldots\ldots\ldots.. (9.25)$$

and$consequently\ M_\theta(t) \longrightarrow \infty\ as\ t \to \infty. A$ similar argument show that $M_\theta(t) \longrightarrow \infty\ as\ t \to \infty$.

$$[M_\theta(t) > e^{tZ'}P_\theta[Z > Z'] = e^{tZ'}v$$

$$where\ P_\theta[Z > Z'] = v > 0, Z' < 0]$$

The $M_\theta(t)$ assume a minimum value at the unique point t^* for which $M'_\theta(t^*) = 0\ now\ M'_\theta(0) = E(Z) \neq 0$, so that $t^* \neq 0$ unless $E_\theta(Z) = 0$.Since$M_\theta(0) = 1\ and\ M_\theta(t^*) < M_\theta(0) = 1$ wherever

$E_\theta(Z) \neq 0$ It must follow that there exist a $t_0 \neq 0$ such that $M_\theta(t_0) = 1$

To prove the condition is necessary, suppose that $P_\theta[Z \geq 0] = 1$ and let$P_\theta[Z = 0] = \alpha < 1$. Thus $P_\theta[Z > 0] = 1 - \alpha$, let $t < 0$ for any $0 < \varepsilon < 1 - \alpha$ we can find positive number C such that

$P_\theta[0 < Z < C] \leq \varepsilon$. Then,

$$\alpha \le M_\theta(t) \le P_\theta[Z=0] + \int_0^C e^{tZ} dF + \int_C^\infty e^{tZ} dF$$

$$= \alpha + \epsilon + e^{tC}(1-\alpha-\epsilon)$$

$$as\ P[Z>C] = 1 - P[Z \le C] = 1 - P[Z=0] - P[0<Z\le C]$$

$\therefore \quad \alpha \le M_\theta(t) \le [\alpha+\epsilon][1-\alpha-\epsilon]e^{tC}$ ……………… (9.26)

And hence, $\quad \alpha \le \lim_{t\to\infty} M_\theta(t) \le \alpha + \epsilon$

Since ε is arbitrary, $\quad \lim_{t\to\infty} M_\theta(t) = \alpha$

We see that, $\quad M'_\theta(t) = \lim_{t\to\infty} \frac{M_\theta(t+h)-M_\theta(t-h)}{h} > 0\ for\ all\ t<0$

and hence $M_\theta(t)=1$ has no solution other than t=0. A similar argument shows that, if $P_\theta[Z\le 0]=1;\ P_\theta[Z=0]<1$ then$M'_\theta(t)<0$, for all t > 0, $M_\theta(t)=1$ has no solution other than t=0. #

<u>Theorem 9.4</u>: [Fundamental Inequality]:

For a given θ and for all t such that $M_\theta(t) > \rho$, where ρ as in Theorem (9.2)

$E_\theta\left[e^{tS_N}\left(M_\theta(t)\right)^{-N}\right] = 1$ …………….. (9.27)

and if $P_\theta[Z>0]>0\ and\ P_\theta[Z<0]>0, where\ Z = log\frac{f(x,\theta_1)}{f(x,\theta_0)}$

$then$ (9.27) $holds\ for\ all\ t.$

<u>Proof</u>: Let the sequential procedure is defined in Theorem 9.2. Then since, $E_\theta e^{tS_n} = E_\theta e^{t(Z_1+\cdots+Z_n)}$

$= \prod_{i=1}^n E_\theta e^{tZ_i} = [M_\theta(t)]^n$ …………..….. (9.28)

$E_\theta[e^{tS_n}[M_\theta(t)]^{-n}] = 1$

$$1 = E_\theta[e^{tS_N}[M_\theta(t)]^{-N}]$$

$$= \sum_{j=1}^{n} P_\theta[N=j]\, E[e^{tS_N}[M_\theta(t)]^{-n}/N=j]$$

$$+ \; P_\theta[N>n]E_\theta[e^{tS_N}[M_\theta(t)]^{-n}/N>n]$$

$= \sum_{j=1}^{n} P_\theta[N=j]\, E\left[e^{tS_j}[M_\theta(t)]^{-j}/N=j\right] + \; P_\theta[N >$
$n]E_\theta[e^{tS_N}[M_\theta(t)]^{-n}/N>n]$ ……………………. (9.29)

Since $E[e^{tS_N}[M_\theta(t)]^{-n}/N=j] = E\left[e^{tS_j}[M_\theta(t)]^{-j}/N=j\right] \, as$

$$\sum_{i=1}^{j} Z_i \text{ is independent of } \sum_{i=j+1}^{n} Z_i$$

Since for N>n, $C_1 < S_n < C_2$ then by (9.29) and Theorem (9.2)

$$0 \le 1 - \sum_{j=1}^{n} P_\theta[N=j]\, E\left[e^{tS_j}[M_\theta(t)]^{-j}/N=j\right] \le \frac{\rho^n}{[M_\theta(t)]^{-n}} E_\theta[e^{tS_n}/N>n]$$

$$= \left(\frac{\rho}{M_\theta(t)}\right)^n k(t)$$

Where k (t) is positive and for fixed θ depends only on t. Letting as $n \to \infty$ we see that for all real t such that $M_\theta(t) > \rho$ equation (9.27) holds.

Suppose now that Z takes on both positive and negative values so that $M_\theta(t)$ has a minimum value which is assumed at t=t* then it follows from (9.29) that for all t,

$P_\theta[N>n] < \frac{[M_\theta(t)]^n}{1<(t)}$ and $P_\theta[N>n] < \frac{[M_\theta(t^*)]^n}{1<(t^*)}$ ……… (9.30)

And hence

$$0 \leq 1 - \sum_{j=1}^{n} P_\theta[N = j]\, E\big[e^{tS_j}[M_\theta(t)]^{-j}/N = j\big] \leq \frac{[M_\theta(t^*)]^n}{1 < (t^*)} \frac{k(t)}{k(t^*)}$$

…………….. (9.31)

Thus $n \to \infty 0 \leq 1 - E_\theta[e^{tS_N}[M_\theta(t)]^{-N}] \leq 0\ as \frac{M_\theta(t^*)}{M_\theta(t)} < 1$

Or $E_\theta[e^{tS_N}[M_\theta(t)]^{-N}] = 1$ #

OC and ASN function of SPRT

For brevity we denote by L (θ) the OC (*operating characteristic function)* of SPRT.

Let us consider the sequence Z_i of independent r.v's defined by $Z_i = log\frac{f(x_i,\theta_1)}{f(x_i,\theta_0)}$ $i = 1,2, \ldots$ satisfying the assumption of theorem (9.2) them if EZ≠0, there exist one and only $h_0 \neq 0$ such that$E(e^{h_0 z}) = 1; if\ E(Z) = 0$, this condition hold only for $h_0 = 0$ let us assume that$E(Z) \neq 0$. Since the distribution of Z depends on θ. Thus let us $h_0 = h_0(\theta)$.

$$M_\theta(h_0) = M(h_0(\theta)) = Ee^{Zh_0(\theta)} = 1 \quad \text{……………. (9.32)}$$

$$= \int e^{Zh_0} f(Z,\theta)\, dZ = 1$$

Or $\sum e^{Zh_0} p(Z,\theta) = 1$ ……………………. (9.33)

$E_\theta e^{S_N h_0(\theta)} = \prod_{i=1}^{N} E\, e^{Z_i h_0(\theta)} = 1$ ………………. (9.34)

$$1 = E_\theta e^{S_N h_0(\theta)} = L(\theta) E_\theta\big(e^{S_N h_0(\theta)}/S_N \leq \log B\big) + 1 - L(\theta) E_\theta\big(e^{S_N h_0(\theta)}/S_N \leq \log A\big)$$

……………. (9.35)

1 = L(θ) E_θ^* + $[1 - L(\theta)]E_\theta^{**}$ ………………(9.36)

Where $E_\theta^*, E_\theta^{**}$ represent the conditional expectations when we accept and reject the hypothesis respectively,

$L(\theta) = \frac{E_\theta^{**}-1}{E_\theta^{**}-E_\theta^{*}}$ ………………… (9.37)

We now find the approximate expression for $L(\theta)$. Let us consider, $S_N = logB\ and\ S_N = logA$ instead of inequality $S_N \leq logB\ and\ S_N \geq logA$. Thus if $S_N = logB$

$$E_\theta^*[exp S_N h_0(\theta)] \approx E_\theta^*[exp(logB)h_0(\theta)]$$

$$\approx E_\theta^*[B]^{h_0(\theta)} \approx [B]^{h_0(\theta)}$$

Similarly, $E_\theta^{**}[exp S_N h_0(\theta)] \approx E_\theta^{**}[exp(logA)h_0(\theta)] \approx [A]^{h_0(\theta)}$

$$\therefore \quad L(\theta) = \frac{[A]^{h_0(\theta)}-1}{[A]^{h_0(\theta)}-[B]^{h_0(\theta)}}$$

When, $E_\theta(Z) = 0, then\ h_0(\theta') = 0$ where θ' is value of θ for which $E_\theta(Z) = 0 = 0$. Then,

$$\lim_{\theta \to \theta'} L(\theta) = L(\theta') = \lim_{\theta \to \theta'} \frac{[A]^{h_0(\theta)}-1}{[A]^{h_0(\theta)}-[B]^{h_0(\theta)}}$$

$$= \lim_{\theta \to \theta'} \frac{\frac{A^{h_0(\theta)}-1}{\theta}}{\frac{A^{h_0(\theta)}-B^{h_0(\theta)}}{\theta}} = \frac{logA}{logA - logB}$$

For any real $h_0(\theta)$, we can determine the point in the plane with co-ordinate (θ, L (θ)). The locus of these points will be approximate graph of the **OCfunction**.

Expected value of N i.e $E_\theta N$ or ASN (*Average Sampling Number*):

We know that for

EZ≠0 $E_\theta[e^{S_N h}[M_\theta(h)]^{-N}] = 1$ differentiating w.r.to h at h=0

$E_\theta\{S_N e^{S_N h}[M_\theta(h)]^{-N} - N e^{S_N h}[M_\theta(h)]^{\overline{-N-1}}(M''_\theta(h))\}_{h=0} = 0$

$$E_\theta\{S_N - NE_\theta Z\} = 0 = E_\theta(N) = \frac{E_\theta(S_N)}{E_\theta(Z)}$$

$E_\theta^*[S_N]$ Denote the conditional expectation of the r.v's provided $S_N \leq logB$ and $E_\theta^{**}[B_N]$ the conditional expectation of S_N provided $S_N \geq logA$.

$$E_\theta(S_N) = L(\theta)E_\theta^*(S_N) + \big(1 - L(\theta)\big)E_\theta^{**}(S_N)$$

$$E_\theta(N) = \frac{L(\theta)E_\theta^* S_N + (1 - L(\theta))E_\theta^{**}(S_N)}{E_\theta(Z)}$$

If $S_N = logB\ orS_N = logA$ according as accepting and rejecting hypothesis.

$$E_\theta(N) = \frac{L(\theta)logB + \big(1 - L(\theta)\big)logA}{E_\theta(Z)}$$

If $E_\theta(Z) = 0$ we differentiate the fundamental Identity twice, we have,

$$E'_\theta\left[\left\{\left(S_N - N\frac{M'_\theta(h)}{M_\theta(h)}\right)^2 - \frac{NM''_\theta(h)M_\theta(h) - N\left(M'_\theta(h)\right)^2}{\left(M_\theta(h)\right)^2}\right\}e^{S_N h}[M_\theta(h)]^{-N}\right] = 0$$

Taking the derivative at h=0 and using

$M_\theta(0) = 1, M'_\theta(0) = E_\theta(Z) = 0$ And $M''_\theta(0) = E_{\theta_1}(Z^2) \neq 0$ we have

$$E_{\theta'}(S_N^2 - NE_{\theta'}Z^2) = 0$$

Or $E_{\theta'}(N) = \frac{E_{\theta'}S_N^2}{E_{\theta'}(Z^2)} = \frac{L(\theta')S_N^2 + (1 - L(\theta'))E_\theta^{**}(S_N^2)}{E_{\theta'}(Z^2)}$

$$= \frac{L(\theta')(logB) + \big(1 - L(\theta')\big)(logA)^2}{E_{\theta'}(Z^2)}$$

$$= \frac{logA}{logA - logB}(logB)^2 + \left(1 - \frac{logA}{logA - logB}\right)(logA)^2 / E_\theta(Z^2)$$

$$= -\frac{logAlogB}{E_{\theta'}(Z^2)}$$

Theorem 9.5: [$wald$] If SPRT is defined by $(logB, logA)$, where

$0 < B < 1, 0 < A < 1$, then the error probabilities α, β satisfy,

$A \leq \frac{1-\beta}{\alpha}, B \geq \frac{\beta}{1-\alpha}$ Where, $\alpha = P\theta_1[S_N \geq A], \beta = P\theta_0[S_N \geq B]$

If we set, $A' = \frac{1-\beta}{\alpha}, B' = \frac{\beta}{1-\alpha}$ then corresponding error probabilities α', β' satisfy, $\alpha' \leq \frac{\alpha}{1-\beta}, \beta' \geq \frac{\beta}{1-\alpha}$, and if $\alpha + \beta \leq 1, then$

$\alpha' + \beta' \leq \alpha + \beta$

Exp 9.1: Let$(X_1, \dots, X_n)$ be i.i.d r.v's having N (θ, 1). The two simple hypotheses are, $H_0: \theta = -1, H_1: \theta = 1$

$$Z = log\frac{f(X,1)}{f(X,-1)} = loge^{-\frac{(x-1)^2}{2}}e^{\frac{(x+1)^2}{2}} = loge^{2x} = 2X$$

m.g.f of X is, $G_\theta^{(t)} = \exp(\frac{t^2}{2} + \theta t)$

m.g.f of 2X is, $M_\theta^{(t)} = e^{2t^2 + 2\theta t}$

It follows that, $h_0(\theta) = -\theta$ thus,

$$L(\theta) = \frac{e^{-\theta a}}{e^{-\theta a} - e^{\theta b}} \ \ where, -b = logB, a = logA$$

$$E_\theta(N) = \frac{1}{2\theta}\left[a\frac{1-e^{\theta b}}{e^{-\theta a} - e^{\theta b}} + b\frac{e^{-\theta a} - 1}{e^{-\theta a} - e^{\theta b}}\right]$$

For$H_0: \theta = \theta_0, H_1: \theta = \theta_1$,

$$\lambda_n = \prod_{i=1}^{n}\frac{f(X_i, \theta_1)}{f(X_i, \theta_0)} \ \ or \ \ log\lambda_n = \sum_{i=1}^{n}\frac{f(X_i, \theta_1)}{f(X_i, \theta_0)} = \sum Z_i$$

$$= \sum_{i=1}^{n}\frac{f(X_i - \theta_1)^2}{2} + \sum_{i=1}^{n}\frac{f(X_i - \theta_0)^2}{2}$$

$$= (\theta_1 - \theta_0)\sum X_i + \frac{(\theta_0^2 - \theta_1^2)n}{2} = \sum Z_i$$

We continue sampling as long as,

$$A < \sum Z_i < B \; or \; \frac{A}{(\theta_1 - \theta_0)} + \frac{(\theta_0^2 - \theta_1^2)n}{2(\theta_1 - \theta_0)} < \sum X_i < \frac{B}{(\theta_1 - \theta_0)} + \frac{n(\theta_0^2 - \theta_1^2)}{2(\theta_1 - \theta_0)}$$

$$Z_1 = (\theta_1 - \theta_0)X_1 + \frac{(\theta_0^2 - \theta_1^2)}{2}$$

$$E_{\theta_i}(Z_1) = (\theta_1 - \theta_0)\theta_i + \frac{(\theta_0^2 - \theta_1^2)}{2} \;\; , i = 0{,}1$$

If $\alpha = .01 \, , \beta = .95$

$$A \approx loga' \; where, a' = \frac{1-\beta}{1-\alpha}$$

$$A \approx loga' = -1.29667$$

$$B \approx logb' = log\frac{\beta}{\alpha} = log\frac{.95}{.01} = log95 = 1.97772$$

$$E_0Z_1 = -\frac{1}{2} = -.5 \;\; , E_1Z_1 = .5$$

$$E_0N \approx \frac{(1-\alpha)A + \alpha B}{E_0Z_1} = \frac{.99(-1.29667) + .01(1.97772)}{-.5} = 2.53$$

$$E_1N \approx \frac{(1-\beta)A + \beta B}{E_1Z_1} = 3.63$$

6.8 Self-Assessment Exercise

1. State and prove the minimax theorem.
2. Explain the role of complete class theorem in estimation theory.
3. Write a note on sequential nature of Bayes theorem and its need.

6.9 Summary

In this unit, section 6.3 and 6.4 discusses the minimax theorem and complete class theorem, respectively. Equalizer rules are covered in section 6.5. The multiple decision problems are discussed in section 6.6. Section6.7 covers the continuous form of Bayes theorem and its sequential nature along with its need.

6.10 Further Readings

1 Lee, P.M. (1997) Bayesian Statistics: An Introduction, Arnold.

2 Leonard, T. and Hsu, J.S.J. (1999) Bayesian Methods, Cambridge University Press.

3 Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods, Springer Verlag.

UNIT-7: BAYESIAN DECISION THEORY

Structure

7.1 Introduction

We encounter lots of decision problems in real life. For example, a mobile store might need to know whether a particular customer based on a certain age, is going to buy a mobile or not. Bayesian Decision Theory helps us in making decisions on whether to select a class with some probability or an opposite class with some other probability based on a certain features. There is always some sort of risk attached to any decision we choose. The entire purpose of the Bayes Decision Theory is to help us select decisions that will cost us the least 'risk'.

7.2 Objectives

After studying this unit, you should be able to describe

- Some basic elements of Bayesian Decision Theory
- Optimal Bays Decision Function
- The Relationship of Bays and Minimax Decision Rules
- The idea of Least Favourable Distribution

7.3 Basic Elements of Bayesian Decision Theory

Mainly there are four elements of Bayesian Decision theory, namely Prior information, Likelihood (rather the joint distribution of the observations), Posterior and risk involved. In the Bayesian framework, we treat the unknown parameter, as a random variable. More specifically, we assume that we have some initial guess about the distribution of this unknown parameter. This distribution is called the prior distribution. After observing some data, we update the distribution of this unknown parameter (based on the prior distribution and thejoint distribution of the observations). This step is usually done using Bayes' theorem. That is why this decision theoretic approach is called the Bayesian decision theory. As there is always some sort of risk attached to any decision we make. The entire purpose of the Bayes Decision Theory is to help us select decisions that will cost us the least 'expected risk' or loss.

7.4 Optimal Bayes Decision Function

Admissibility is a useful criterion when searching for optimal decision rules as the optimal decision rule gives the minimum error. For example, knowing that an estimator is inadmissible is clearly bad in that another estimator with lower risk is guaranteed to exist. One of the most popular examples of an inadmissible estimator is given by James and Stein (1961). A detailed discussion on the optimality is already given in section 2.4 and 2.5 of Block 1.

7.5 Relationship of Bayes and Minimax Decision Rule

This section explores some interesting results to develop an understanding about the relationship between Bayes and minimax decision rules. Minimax is a decision rule used in decision theory, game theory, statistics, etc for minimizing the possible loss for a maximum loss scenario. When dealing with gains, it is referred to as "maximin" – to maximize the minimum gain. Hence, in this approach one tries to guard against the highest possible risk in a pessimist's way i.e. by trying to keep the smallest of the highest possible risks. This can be proved that such a rule always exists. Whereas a Bayes rule is the decision rule in the class of decision rules that has the smallest average risk. Hence it is obvious that if the Bayes rule has constant risk, then it is minimax.

7.6 Least Favourable Distributions

Let for some decision problem ,δ_1 and δ_2 be two two Bayes rules w.r.t. prior distributions g_1 and g_2 , respectively. Then, g_1 is called least favourable prior distribution if $r(g_1,\delta_1) \geq r(g_2,\delta_2)$ irrespective of g_2.

7.7 Self-Assessment Exercise

1. If there exist a prior g for some unknown parameter say, μ and let δ_g be a Bayes rule corresponding to g and if $r(g,\delta_g) \geq \sup_\mu r(\mu,\delta_g)$; then (i) δ_g is a minimax rule, (ii) g is the least favourable prior distribution.

2. Define the concept of optimal Bayes decision functions.

7.8 Summary

In section 7.3, some basic elements of Bayesian decision theory have been discussed. Section 7.4 discusses about the optimality criteria for decision functions. Section 7.5 explores the relationship between Bayes and Minimax Decision Rules. Then, section 7.6 defines the Least Favourable Distribution.

7.10 Further Readings

1. Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods, Springer Verlag.
2. Berger, J.O. (1985). Statistical decision theory-Fundamental concepts and methods, Springer Verlag.

UNIT-8: BAYESIAN INTERVAL ESTIMATION

Structure

8.1 Introduction

Estimation is used to come to some conclusions regarding an unknown population parameter with the help of a sufficiently large sample from that population. Having obtained the estimate of unknown parameter from a given sample, the problem is, "Can we make some reasonable probability statements about the unknown parameter a in the population, from which the sample has been drawn". To answer such questions, we use the technique of Interval estimation. Classical approach covers such problems in confidence interval estimation whereas in modern or subjective approach Bayesian interval estimation covers such problems.

8.2 Objectives

After studying this unit, you should be able to

- Define the concept of sufficiency in Bayesian sense

- Explore the use of different priors.
- Test the hypothesis in Bayesian's way
- Elaborate the empirical Bayes Procedures.

8.3 Bayesian Sufficiency

Kolmogorov, Raifa Scefferetc have discussed various statistical concepts from Bayesian point of view in detail. But here we will discuss the concept of sufficiency first in classical sense and then in Bayesian sense. Consider, (X,ζ) is a measurable space carrying a family of probability measures on parametric space Θ. Then, classical sufficiency is defined as the conditional probability on ζ given any sub σ-field is independent of parameter in Θ, but in Bayesian sense given any prior ξ on (Θ, A), the posterior on Θ is the same as ζ st$A$ is a σ-field. Because of the compelling reasons to perform a conditional analysis and the alternatives of using Bayesian machinery to do so there have been attempts to use the Bayesian approach even when no (or minimal) prior information is available. What is needed in such situation is a Non informative prior, by which is meant a prior which contains no information about θ (or more crudely which 'faros' no possible values of θ over others.) for example, in testing between two simple hypothesis, the prior which gives probability ½ to each of the hypothesis is clearly non-informative.

Exp: suppose the parameter of interest is normal mean θ, so that the parameter space$\Theta = \{-\infty, \infty\}$. If non-informative prior density is desired, it seems reasonable to give equal weights to all possible values of θ. unfortunately, if $\pi(\theta) = c > 0$ is chosen, the π has infinite mean $i.e \int \pi(\theta)d\theta = \infty$ and is not proper density. Nevertheless, such π can be successfully worked with the choice of c is unimportant, so that typically the non-informative prior clearly for this problem is chosen to be π(θ)=1 this is often called the informative density on R and was intersected and used by Laplace (1812).

As in the above example, it will frequently happen that natural non-informative prior is an improper prior, namely which has infinite mass.

Exp: instead of considering θ, suppose the problem has been parameterized in terms of$\eta = e^{\theta}$, this is one-to-one information and should have no bearing on the ultimate answer.

But if π (θ) is the density of θ, then the correspondently for η is,

$\pi^{*}(\eta) = \eta^{-1}\pi(log\eta)$ Hence if the non-informative prior of θ is chosen to be constant, we should choose the non-informative prior of η to be conditional to η^{-1} to maintain consistency. Thus we maintain consistency and choose both the non-informative prior

Non informative Priors for location and scale parameters:

Exp: suppose that $\mathfrak{x}$ and Θ are subsets of R^k, and that the density of $\underline{X}$ is of the form $f(\underline{x} - \underline{\theta})$ $i.e$ depend on $(\underline{x} - \underline{\theta})$. The density then said to be a location density, and θ is called a location parameter. (Sometimes a location vector when $k \geq 2$). The $N(\theta, \sigma^2), \sigma^2$ fixed, is an example of location density.

To derive a non-informative prior for this situation, imagine that, insisted of observing X, we observe the random variable $\underline{Y}$=$\underline{X}$+$\underline{C}$. CϵR^k . Define $\underline{\eta} = \underline{\theta} + \underline{C}$ it is clear that Y has density $f(\underline{y} - \eta)$. If now

$\mathfrak{x} = \Theta = R^k$ Thus the sample space and parameter space for (Y, η) problem are also R^k. The (X, Θ) &(Y, η) problems are identical and sensitive and it seems reasonable to in sets that they have the same non-informative prior.

Letting π and π* denote the non-informative priors in the (X, Θ) and (Y, η) problems respectively, the above arguments implies that π and π* should be equal i.e

$$p^{\pi}[\theta\epsilon A] = p^{\pi^*}[\eta\epsilon A]$$

For any set A in R^k. Since η=θ+C, it should be true that

$$p^{\pi^*}[\eta\epsilon A] = p^{\pi}[\theta + C\epsilon A] = p^{\pi}[\theta\epsilon A - C]$$

$A - C = \{Z - C : Z\epsilon A\}$ $then,$

$p^{\pi}[\theta\epsilon A] = p^{\pi}[\theta\epsilon A - C]$ $for\ all\ \theta\epsilon R^k$ …………………… (1)

Any π satisfying relation (1) is said to be location invariant prior.

Assuming that the prior has a density then,

$$\int_A \pi(\theta)d\theta = \int_{A-C} \pi(\theta)d\theta = \int_A \pi(\theta - C)d\theta \qquad for\ all\ A\epsilon R^k$$

$$\pi(\theta) = \pi(\theta - C) \quad for\ all\ \theta\epsilon\Theta, or\ \pi(C) = \pi(0) \quad for\ all\ C\epsilon R^k$$

This conclusion is that π must be constant function. It convenient to choose the constant to be 1, so the non-informative prior density for a location parameter is $\pi(\theta) = 1$

A one-dimensional scale density is a density of the form, $\alpha^{-1} f(\frac{x}{\alpha})$ where$\alpha > 0$. The parameter $\alpha > 0$ is called a scale parameter. The

$N(0, \sigma^2) G(\alpha, \beta), \alpha$known as scale density.

To derive a non-informative prior for this situation, imagine that, instead of observing X, we observe the random variable Y=CX C$>$ 0.

Define $\eta = C\alpha$, can easy calculation show that the density of Y is

$\eta^{-1} f(\frac{y}{\eta})$. If 𝔵=R or (0,∞) then the sample and parameter space for the (X, α) problems are the same as there for the (Y, η) problem. The two problems are thus identical in structure, which again indicates that they should have the same non-informative prior. Letting π and π* denote the priors in the (X, α) and (Y, η) problem, respectively, this means that the equality,

$$p^{\pi}[\alpha\epsilon A] = p^{\pi^*}[\eta\epsilon A]$$

$Should\ for\ all\ A \subset (0, \infty)$. Since η= $C\alpha$, it should also be true that

$$p^{\pi^*}[\eta\epsilon A] = p^{\pi}[\alpha\epsilon C^{-1}A],$$

$C^{-1}A = \{C^{-1}Z : Z\epsilon A\}$. Putting these together, it follows that π should satisfy,

$$p^{\pi}[\alpha\epsilon A] = p^{\pi}[\alpha\epsilon C^{-1}A] \qquad for\ all\ C > 0$$

And any distribution π for which this is true is called scale invariant.

$$\int_A \pi(\alpha)d\alpha = \int_{C^{-1}A} \pi(\alpha)d\alpha$$

$$= \int_A \pi(C^{-1}\alpha)C^{-1}d\alpha \qquad for\ all\ A \subset (0,\infty) \Rightarrow \pi(\alpha)$$

$$= C^{-1}\pi\,(C^{-1}\alpha) \qquad for\ all\ \alpha.\ let\ \alpha = C$$

$\pi(C) = C^{-1}\pi$ (1). Setting for convenience, and nothing that above equality must hold for all$C > 0$, it follows that a reasonable non-informative for a scale parameter is $\pi\alpha = \alpha^{-1}$.

Non-informative prior in general setting:

For more general problem, various (somewhat ad hoe) suggestive have been advance for determining a non-informative prior. The most widely used method is that of Jeffrey's method which is as follows:

If $\underline{\theta} = (\theta_1, \dots, \theta_k)'$ is a vector, Jeffrey's suggest the use of

$$\pi(\underline{\theta}) = [\det I(\underline{\theta})]^{\frac{1}{2}} \qquad 'det' = determinant;$$

Where $I(\underline{\theta}) = [I_{ij}(\underline{\theta})] \Rightarrow I_{ij}(\underline{\theta}) = -E_{\underline{\theta}}[\frac{\partial^2}{\partial\theta_i\partial\theta_j} logf(x/\underline{\theta})]$

Exp: A location-scale density is a density of the form $\sigma^{-1}f(\frac{x-\theta}{\sigma})$ where $\theta \epsilon R, \sigma > 0$ are the unknown parameters. $N(\theta,\sigma^2)$is crucial example of location-scale density Working with$N(\theta,\sigma^2)$, $\underline{\theta}$= (θ, σ). Fisher informative matrix is,

$$I(\underline{\theta}) = -E_{\underline{\theta}} \begin{pmatrix} \frac{\partial^2}{\partial\theta^2}\frac{(x-\theta)^2}{2\sigma^2} & \frac{\partial^2}{\partial\theta\,\partial\sigma}\frac{(x-\theta)^2}{2\sigma^2} \\ \frac{\partial^2}{\partial\theta\,\partial\sigma}\frac{(x-\theta)^2}{2\sigma^2} & \frac{\partial^2}{\partial\theta^2}\frac{(-(x-\theta)^2)}{2\sigma^2} \end{pmatrix}$$

$$= -E_{\underline{\theta}} \begin{pmatrix} \frac{-1}{\sigma^2} & \frac{2(\theta-x)}{\sigma^3} \\ \frac{2(\theta-x)}{\sigma^3} & \frac{-3(x-\theta)^2}{\sigma^4} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{3}{\sigma^2} \end{pmatrix}$$

$$\pi(\underline{\theta}) = [\frac{1}{\sigma^2}, \frac{3}{\sigma^2}]^{\frac{1}{2}} \alpha \frac{1}{\sigma^2}$$

This is actually the non-informative prior ultimately recommended by Jeffrey's non-informative prior is that it is not affected by restriction on the parameter space. Thus if it is known that $\Theta > 0$, the Jeffrey's non-informative prior is still $\pi(\theta) = 1$.

Exp: let $(X_1, \dots, X_n)$ be a random sample from N (θ_1, θ_2) let the non-informative prior of (θ_1, θ_2) be $(\theta_1, \theta_2) \propto \frac{1}{\theta^2}$ and $\theta_1 \& \theta_2$ assumed to be independent. Find the posterior $.d.f\ of f(\theta_1/\underline{x}) \& f(\theta_2/\underline{x})$.

Solution: $f(x_1, \dots, x_n/\theta_1, \theta_2) \propto \frac{1}{(\theta_2)^{\frac{n}{2}}} \exp -\frac{\sum(x_i - \theta_1)^2}{2\theta^2}$

$$f(\theta_1, \theta_2/x_1, \dots, x_n) \propto \frac{1}{(\theta_2)^{\frac{n}{2}}} \exp -\frac{\sum(x_i - \theta_1)^2}{2\theta^2} \frac{1}{\theta^2}$$

$$= \frac{1}{(\theta_2)^{\frac{n}{2}} \theta_2} \exp -\frac{\sum(\bar{x} - \theta_1)^2}{2\theta^2} \exp -\frac{n(\bar{x} - \theta_1)^2}{2\theta^2}$$

$$= \frac{1}{(\theta_2)^{\frac{n+2}{2}}} \exp -\frac{S^2 n - 1}{2\theta^2} \exp -\frac{n(\bar{x} - \theta_1)^2}{2\theta^2}$$

$f(\theta_1/\underline{x}) \propto \int_0^\infty \frac{1}{(\theta_2)^{\frac{n+2}{2}}} \exp -\frac{\sum(x_i - \theta_1)^2}{2\theta^2} d\theta_2$ Put $\frac{1}{2\theta^2} = t \Rightarrow -\frac{d\theta_2}{\theta_2^2} = 2dt$

$$\propto \int_0^\infty t^{\frac{n+2}{2}} \exp -\sum(x_i - \theta_1)^2 t \, \frac{1}{t} \, dt$$

$$= \int_0^\infty t^{\frac{n}{2}-1} \exp -t \sum(x_i - \theta_1)^2 \, dt$$

$$\propto \frac{1}{[\sum(x_i - \theta_1)^2]^{\frac{n}{2}}} = \frac{1}{[\sum(x_i - \bar{x})^2 + n(\bar{x} - \theta_1)^2]^{\frac{n}{2}}}$$

$$\propto \frac{1}{[1+\frac{n(\bar{x}-\theta_1)^2}{\Sigma(x_i-\bar{x})^2}]^{\frac{n}{2}}} = \frac{1}{\left[1+\frac{T^2}{n-1}\right]^{\frac{n-1}{2}}}$$

$$Where, T \sim t - distribution\ with\ (n-1\ ) degree\ of\ freedom.$$

$$f(\theta_2/\underline{x}) \propto \frac{1}{(\theta_2)^{\frac{n+2}{2}}} \exp - \frac{\overline{n-1}s^2}{2\theta^2} \int_{-\infty}^{\infty} \exp - \frac{n(\bar{x}-\theta_1)^2}{2\theta^2} d\theta_1$$

$$\propto \frac{(\theta_2)^{\frac{1}{2}}}{(\theta_2)^{\frac{n+2}{2}}} \exp - \frac{\overline{n-1}s^2}{2\theta^2}$$

$$= \frac{1}{(\theta_2)^{\frac{n+1}{2}}} \exp - \frac{\overline{n-1}s^2}{2\theta^2}$$

Let $w = \frac{\overline{n-1}s^2}{\theta^2}$ $\quad$ $dw = \frac{-\overline{n-1}s^2}{\theta_2^2}\ d\theta_2$

$$f(w/\underline{x}) \propto \frac{(\theta_2)^{\frac{1}{2}}}{(\theta_2)^{\frac{n+1}{2}}} \exp - \frac{w}{2} = \frac{1}{(\theta_2)^{\frac{n-3}{2}}} \exp - \frac{w}{2}$$

$$= \frac{1}{(\theta_2)^{\frac{n-1}{2}-1}} \exp - \frac{w}{2} \propto \chi^2_{n-1}$$

8.4 Improper Prior Densities

After a detailed discussion in preceding section, it is very much clear that in Bayesian procedures, we update the observed information with the help of prior information called prior densities. But sometimes this information is not integrable or does not have a finite integral, but we as statistician has to make use of this. Such prior densities are termed as improper prior densities. Examples of improper priors include: The uniform distribution on an infinite interval (i.e., a half-line or the entire real line). The beta distribution for α=0, β=0.

8.5 Natural Conjugate Bayesian Density

The concept, of **Natural Conjugate Bayesian Density** or conjugate prior, was introduced by Howard Raiffa and Robert Schlaifer in their work on Bayesian decision theory.A similar concept had been discovered independently by George Alfred Barnard.

In Bayesian probability theory, if the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior and posterior are then called **conjugate distributions,** and the prior is called a **conjugate prior**. For example, beta prior is a conjugate prior for a binomial population. Similarly, gamma is for Poisson population.

8.6 HPD Regions and Bayesian Inference for Normal Populations

For this topic, please refer to section 4.5 of Block 1.

8.7 Empirical Bayes Procedures

The purpose here is to give a simple introduction to empirical Bayes methods. **Empirical Bayes methods** are the procedures in which the prior probability distribution is estimated from the data itself. Thus, this approach stands in contrast to standard Bayesian methods, for which the prior distribution is fixed before any data are observed. Empirical Bayes methods have been around for quite a long time. Their roots can be traced back to work by von Mises in the 1940's, but the first major work must be attributed to Robbins (1955). These procedures further can be classified into "parametric empirical Bayes procedures" and "non-parametric empirical Bayes procedures". The major difference is that the parametric approach specifies a parametric family of prior distributions, while the non-parametric approach leaves the prior completely unspecified. For example, if n iid observations are taken from $f_\lambda(.)$ and the prior distribution for the parameter λ is g(.), then the empirical Bayes estimate of parameter λ using the posterior mean is

$$E[\lambda \mid x_n]=(x_n+1)\ m(x_n+1)/m(x_n)\ \text{(m(.) is the marginal distribution of } X_{i=1,2,3,\ldots,n})$$

$$=(x_n+1)(\text{number of } x_i \text{ equal to } (x_n+1))/(\text{number of } x_i \text{ equal to } x_n)$$

In particular, if the sample is (0,4,2,8,7,4,0,9,3), then n^{th} observation is 3 then the empirical Bayes estimate of parameter λ is (3+1)(2)/(1)=8.

8.8 Posterior Odd Ratio and Bayesian Testing of Hypothesis

Let an event A occurs with probability P[A], then the ratio P[A]/(1-P[A]) is called odds in favour of A (say O[A]) and (1-P[A])/P[A] is called odds against A. Hence, in usual notations, using Bayes theorem, we get $O(H_0 | x) = P(H_0 | x)/ P(H_1 | x)$ called posterior odds on H_0. Which gives $O(H_0 | x) = O(H_0)\ P(x | H_0)/ P(x | H_1)$ i.e. $O(H_0 | x)/ O(H_0) = P(x | H_0)/ P(x | H_1)$ called the Bayes Factor in favour of H_0 (say B_{01}) which is the ratio of two conditional probabilities of data in hand. Jeffreys recommended the following table for testing of hypothesis using Bayes Factors:

Value of $\mathrm{Log}_{10}(B_{10})$	Description
0-0.5	Not substantial evidence against H_0
0.5-1	Substantial evidence against H_0
1-2	Strong evidence against H_0
>2	Decisive evidence against H_0

8.9 Self-Assessment Exercise

1. Explain the concept of Bayes factor and its role in statistical inference.

2. Test H_0: $\lambda=2$ against H_1: $\lambda\neq2$ using single observation from Pois(λ) st λ is a Gamma (2,3) variate.

8.10 Summary

This unit starts with a detailed discussion over Bayesian Sufficiency and Improper Prior Densities, then section 8.5 further explores Natural Conjugate Bayesian Densities. Next then it covers HPD Regions and Bayesian Inference for Normal Populations. Then a bit of Empirical

Bayes Procedures and Posterior Odd Ratio along with their use in Bayesian Testing of Hypothesis is discussed at the end.

8.11 Further Readings

1. Bernardo, J.; Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley.
2. Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
3. *Lee, P. M. (2012). Bayesian Statistics: an introduction. Wiley.*
4. *Winkler, Robert (2003). Introduction to Bayesian Inference and Decision (2nd ed.). Probabilistic.*

U.P.RajarshiTandon Open University, Prayagraj

MScSTAT – 301N /MASTAT – 301N Decision Theory & Bayesian Analysis

Block: 3 Bayesian Analysis

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Pramendra Singh Pundir **Writer**
Department of Statistics
University of Allahabad, Prayagraj

Prof. G. S. Pandey (Rtd.) **Editor**
Department of Statistics
University of Allahabad, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 301N/ MASTAT – 301N DECISION THEORY & BAYESIAN ANALYSIS

First Edition: ***July 2023***
ISBN :

Printed By:

Block & Unit Introduction

The present block of this SLM has three units.

The ***Block - 3 – Bayesian Analysis*** has three units. This block comprises

Unit – 9 – Prior and Posterior Distributions, comprises the A detailed note on prior and posterior distributions.

In ***Unit – 10 – Bayesian Inference Procedures***, we have discussed the theory of Bayesian Inferential procedures.

Unit – 11 – Bayesian Robustness, gives the idea of Bayesian robustness.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT-9: PRIOR AND POSTERIOR DISTRIBUTIONS

Structure

9.1 Introduction

In Bayesian theory, a very important concept is of Subjective probability. It is a type of probability derived from an individual's personal judgment or own experience about whether a specific outcome is likely to occur. It may or may not contain any formal calculations; hence generally it only reflects the subject's opinions and past experience. Thus, subjective probabilities differ from person to person and contain a high degree of personal bias. In Bayesian context it plays an important role as here the theory makes use of posterior density which highly depends on the prior. In this unit different types of priors have been discussed.

9.2 Objectives

After studying this unit, you should be able to

- Define the concept of subjectivity
- Choose a suitable prior for different cases
- Obtain the conjugate prior

9.3 Subjective Probability its Existence and Interpretation

The world is an uncertain place, and the outcome of future events is mostly unpredictable. But we always try to become surer about the future. For this we need information about the event of interest that is about to occur in future like it may rain tomorrow or it may not; you might be hired after a job interview, or you might not. Many scenarios are simply too complex to describe even theoretically and do not allow for repeated experimentation that could be used to assess the chances favouring them. So, here we work with our own belief which may or may not be based on some facts. And such an estimate of the likelihood of an event is called subjective probability, which may be the only option available in such cases. Thus, subjective probability is determining the likelihood of an event based on one's opinion or belief and not on any observations or calculations.

9.4 Subjective Determination of Prior and Posterior Distribution

There are always 50%-50% chances that the fair coin will land with a head and tail up, but one can predict the output of flipping a coin on the basis of one’s belief. For example, one may decide that the distribution in some condition is 60%-40%. This will work as the prior distribution for Bayesian analysis in this case. And this belief gets updated in presence of observations then the updated distribution is called the posterior distribution. In this case, this may become 55%-45% after updation using Bayes theorem.

9.5 Improper Priors, Non-Informative Priors, Invariant Priors

Most of the times, these priors are based on one’s belief hence they may not hold the form of some distribution and hence become improper. Mathematically, their integral does not equals unity. Such priors are called improper priors (as discussed earlier in block 1). These priors may be lead to badly behaved posteriors and paradoxes.

In another situation, if the experimenter does not have any prior information or idea about the distribution of the unknown parameter, then the prior that represents this situation of complete initial ignorance is called a non-informative prior. In such situations, one may refer to the suggestion of Laplace that take uniform distribution as prior in absence of sufficient reason for assigning unequal probabilities to the values of the unknown parameter in the parametric space. A variety of such rules have been proposed but two of the most popular rules are first due to Laplace (discussed earlier) and second-one is due to H. Jeffrey. Jeffrey suggested a thumb rule for determining a non-informative prior for a scale parameter (say μ) as follows:

Rule 1: If $\mu \epsilon [a,b]$, where a and b are finite or infinite then take the prior $g(\mu)$=constant.

Rule 2: If $\mu \epsilon (0,\infty)$, assume (log μ) to be uniformly distributed over the whole real line and take $g(\mu) \propto 1/\mu$.

Here, if μ is replaced with any linear transformation $\lambda = c\,\mu + d$ for any choice of $c(\neq 0)$ and d; then rule 1 suggests the non-informative prior $g(\lambda)$=constant i.e. rule 1 is invariant with respect to linear transformations, similarly rule 2 is invariant under exponential transformation $\lambda = \mu^{k}$ st $k \neq 0$.

9.6 Conjugate Prior Families and Their Construction

In addition to the discussion on conjugate priors in preceding blocks, here we will learn more about the conjugate priors. These priors are sometimes called objective priors because the sampling distribution completely determines the class of prior distributions.

Here we will learn a thumb rule for constructing a conjugate prior. Suppose t(x) is a sufficient statistic for the parameter μ. Then, using Neyman factorization theorem we can write the likelihood as $L(\underline{x},\mu)=k(t(\underline{x}),\mu)h(\underline{x})$ st $\underline{x}=(x_1,x_2,...,x_n)$ and $k(t(\underline{x}))$ is the kernel of likelihood. Replace all the terms that are functions of sample in the kernel, by prior hyperparameters to get the conjugate prior.

Example: Let $(x_1,x_2,...,x_n)$ be a sample from Gamma(m,μ) with m>0 known, giving the kernel to be $k(t(\underline{x}),\mu) = \mu^{-nm} \exp(-t/\mu)$. Therefore, the respective conjugate prior is

$g(\mu)=c\mu^{-a} \exp(b/\mu)$, which is inverted gamma (a-1, b) with hyperparameters ‘a’ and ‘b’.

9.7 Self-Assessment Exercise

1. Prepare a list of conjugate prior families in different cases and verify.

2. Explain the concepts of Improper Priors, Non-Informative Priors, Invariant Priors along with their merits and demerits.

3. Explain the concept of subjectivity and explain the related issues.

9.8 Summary

This Unit covers some very interesting and important concepts of Bayesian approach like subjectivity, Improper Priors, Non-Informative Priors, Invariant Priors and conjugate prior families. Also, the thumb rule for constructing a conjugate prior for given case equips the learner to handle the situation in a relatively more mathematically tractable way.

9.9 Further Readings

1. Berger, J.O. (1993) Statistical Decision Theory and Bayesian Analysis, Springer Verlag.
2. Bernando, J.M. and Smith, A.F.M. (1994). Bayesian Theory, John Wiley and Sons.
3. Box, G.P. and Tiao, G.C. (1992). Bayesian Inference in Statistical Analysis, Addison-Wesley.
4. Leonard, T. and Hsu, J.S.J. (1999) Bayesian Methods, Cambridge University Press.
5. Robert, C.P. (1994). The Bayesian Choice: A Decision Theoretic Motivation, Springer.

UNIT-10: BAYESIAN INFERENCE PROCEDURES

Structure

10.1 Introduction

The Bayesian approach to inference usually refers to prior, posterior, and predictive distributions to obtain estimates of unknown parameters, compare models and test hypotheses. Bayesian methods are now becoming widely accepted as a way to solve applied problems of real world. In this unit a few aspects of Bayesian inference are discussed to equip the learners with some basic understanding of these topics.

10.2 Objectives

After studying this unit, you should be able to

- Explain the Bayesian approach to inference
- Define Credible sets
- Perform testing of hypothesis in Bayesian sense
- Define Generalized Bayes Procedures, Admissibility and minimaxity of Bayes

10.3 Bayesian Inference

Bayesian inference techniques specify how one should update one's beliefs upon observing data. Bayesian updating is particularly important in the dynamic analysis of a sequence of data. Thus, Bayesian inference plays an important role in statistics. Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, sports etc. More detailed theory of Bayesian Inferential procedures and examples are given in Block 1 and 2.

10.4 Credible Sets

We have now learnt that Bayesian credible intervals incorporate problem-specific contextual information from the prior information and in Bayesian analysis it is of interest to find the optimal set, i.e. the smallest set with posterior probability at least, with respect to each prior in the class, called a credible set. Thus, Bayesian credible sets can be treated as the correct name for Bayesian "confidence intervals" (discussed earlier). More specifically, if any set $A \epsilon \Theta$, wrt a posterior $\pi(\theta|x)$ has the credible probability $P(\theta \in A|x) = \int_A \pi(\theta|x) d\theta$, then A is called a credible set for θ.

10.5 Testing of Hypothesis

This topic has already been covered under the topic "Posterior Odd Ratio and Bayesian Testing of Hypothesis" in detail in Block 2.

10.6 Generalized Bayes Procedures, Admissibility and Minimaxity of Bayes

These topics have already been covered in detail in Block 1 and Block 2.

10.7 Self-Assessment Exercise

1. Define the concept of credible sets and their role in inference.

2. Define the relationship between credible sets and testing process.

10.8 Summary

Though most of the topics in this unit have already been covered but still this unit gives a sight to explore those topics in the light of credible sets.

10.9 Further Readings

1. Gemerman, D and Lopes, H. F. (2006) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman Hall.
2. Lee, P.M. (1997) Bayesian Statistics: An Introduction, Arnold.
3. Leonard, T. and Hsu, J.S.J. (1999) Bayesian Methods, Cambridge University Press.
4. Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods, Springer Verlag.

Unit-11: Bayesian Robustness

Structure

11.1 Introduction

Bayesian analysis, also called Bayesian sensitivity analysis, is a type of sensitivity analysis applied to the outcome from Bayesian inference or Bayesian optimal decisions. Robust Bayesian analysis, also called Bayesian sensitivity analysis, investigates the robustness of answers from a Bayesian analysis to uncertainty about the precise details of the analysis. Robust Bayes methods acknowledge that it is sometimes very difficult to come up with precise distributions to be used as priors. Likewise, the appropriate likelihood function that should be used for a particular problem may also be in doubt. In a robust Bayes approach, a standard Bayesian analysis is applied to all possible combinations of prior distributions and likelihood functions selected from classes of priors and likelihoods considered empirically plausible by the analyst. In this approach, a class of priors and a class of likelihoods together imply a class of posteriors by pair-wise combination through Bayes rule.

11.2 Objectives

After studying this unit, you should be able to

- Define the idea of Bayesian Robustness.
- Define Markov Chain Monte Carlo (MCMC) techniques.

- List the methods involved in Monte Carlo integration.

11.3 Ideas of Bayesian Robustness

Broadly robustness defines the sensitivity of the estimates. Bayesian analysis, also called Bayesian sensitivity analysis, is a type of sensitivity analysis applied to the outcome from Bayesian inference or Bayesian optimal decisions. Robust Bayesian analysis, also called Bayesian sensitivity analysis, investigates the robustness of answers from a Bayesian analysis to uncertainty about the precise details of the analysis. Robust Bayes methods acknowledge that it is sometimes very difficult to come up with precise distributions to be used as priors. Likewise the appropriate likelihood function that should be used for a particular problem may also be in doubt. In a robust Bayes approach, a standard Bayesian analysis is applied to all possible combinations of prior distributions and likelihood functions selected from classes of priors and likelihoods considered empirically plausible by the analyst. In this approach, a class of priors and a class of likelihoods together imply a class of posteriors by pair-wise combination through Bayes rule. Robust Bayes also uses a similar strategy to combine a class of probability models with a class of utility functions to infer a class of decisions, any of which might be the answer given the uncertainty about best probability model and utility function. In both cases, the result is said to be robust if it is approximately the same for each such pair. If the answers differ substantially, then their range is taken as an expression of how much (or how little) can be confidently inferred from the analysis.

11.4 Asymptotic Expansion for Posterior Density

A framework for Bayesian inference: - Additional information which may update beliefs about $\theta$are usually in the form of observed data $x_1, x_2, \ldots, x_n$. The information regarding θ contained in the data is represented by the likelihood function. Bayes' theorem can also be used to update beliefs about a parameter θ after data are observed. The updated beliefs are represented by the posterior distribution. The posterior distribution, which summarizes all the information available about θ after observing data, is the primary focus of Bayesian inference.

Beliefs about an unknown parameter θ are also represented probabilistically in Bayesian statistics. A subjective estimate can be made of the probability that the value of θ is θ_1, say, that is, of the probability $P(\theta = \theta_1)$, for some value θ_1.

If you are certain that $\theta = \theta_1$, then $P(\theta = \theta_1) = 1$. However, the value of θ is rarely known with certainty. Instead, there will be other values of θ that are possible. Usually, the possible values of θ are all values in some continuous interval. For example, if θ is a proportion, then the true value of θ could potentially be any value in the interval $[0, 1]$. However, for simplicity, first suppose that θ can only be one of a set of discrete values $\theta_1, \theta_2, \ldots, \theta_n$. For each possible value θ_i, the probability $P(\theta = \theta_i)$ can be estimated subjectively, so that $P(\theta = \theta_i)$ represents beliefs about whether or not $\theta = \theta_i$. If $P(\theta = \theta_i)$ is estimated for all possible values of θ_i, then these probabilities will form a probability distribution for θ. This probability distribution gives a probabilistic representation of all the available knowledge about the parameter θ, and is known as the prior distribution, or simply the prior.

Suppose that the random variable X has some distribution with unknown parameter θ. If it were known that the value of $\theta$is θ_0, then the distribution of X would be known exactly. If X is discrete then, conditional on $\theta = \theta_0$, the (conditional) probability mass function $p(x|\theta = \theta_0)$ can be written down. Similarly, if X is continuous, the conditional probability density function $f(x|\theta = \theta_0)$ can be written down.

Given an observation x on a discrete random variable X, the value of the conditional p.m.f. $p(x|\theta = \theta_0)$ can be calculated for each possible value θ_0 of θ. Since a value is defined for each possible value of θ, these values can be viewed as values of a function of θ, which can be written $p(x|\theta)$. This function is called the likelihood function, or simply the likelihood. It represents how likely the possible values of θ are for the observed data x.

More generally, in a statistical inference problem, the data consist of n independent observations $x_1, \ldots, x_n$on X. In this case, the likelihood is of the following form:

$L(\theta) = p(\text{data}|\theta) = p(x_1|\theta) \times \cdots \times p(x_n|\theta)$ifX is discrete,

$L(\theta) = f(\text{data}|\theta) = f(x_1|\theta) \times \cdots \times f(x_n|\theta)$ifX is continuous.

11.5 Bayesian Calculation

Suppose a 30-year-old man has a positive blood test for a prostate cancer marker (PSA). Assume this test is also approximately 90% accurate. In this situation, the individual would like to know the probability that he has prostate cancer, given the positive test, but the information at hand is simply the probability of testing positive if he has prostate cancer, coupled with the knowledge that he tested positive. Bayes theorem offers a way to reverse conditional probabilities and, hence, provides a way to answer these questions.

Bayesian probability is one of the major theoretical and practical frameworks for reasoning and decision making under uncertainty. The historical roots of this theory lie in the late 18th, early 19th century, with Thomas Bayes and Pierre-Simon de Laplace.

In its raw form, Bayes Theorem is a result in conditional probability, stating that for two random quantities yand θ,

$$p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)},$$

wherep(·) denotes a probability distribution, and p(· | ·) a conditional distribution. Where y represents data and θ represents parameters in a statistical model, Bayes Theorem provides the basis for Bayesian inference. The 'prior' distribution p(θ) (epistemological uncertainty) is combined with 'likelihood' p(y|θ) to provide a 'posterior' distribution p(θ|y) (updated epistemological uncertainty): the likelihood is derived from an aleatory sampling model p(y|θ) but considered as function of θ for fixed y.

11.6 Monto Carlo Integration

Monte Carlo methods are numerical techniques which rely on random sampling to approximate their results. Monte Carlo integration applies this process to the numerical estimation of integrals. Monte Carlo integration uses random sampling of a function to numerically compute an estimate of its integral.

11.7 Markov Chain Monto Carlo Techniques

Markov Chain Monte Carlo (MCMC) techniques are methods for sampling from probability distributions using Markov chains. MCMC methods are used in data modeling for Bayesian inference and numerical integration. MCMC techniques aim to construct cleverly sampled chains which draw samples which are progressively more likely realizations of the distribution of interest. Here, Monte Carlo methods are numerical techniques which rely on random sampling to approximate their results. Monte Carlo integration applies this process to the numerical estimation of integrals. Monte Carlo integration uses random sampling of a function to numerically compute an estimate of its integral. Suppose that we want to integrate the one-dimensional function f(x) from ato b:

$$F = \int_a^b f(x)dx$$

We can approximate this integral by averaging samples of the function f at uniform random points within the interval. Given a set of N uniform random variables $X_i \in [a, b)$ with a corresponding pdf of $1/(b - a)$, the Monte Carlo estimator for computing F is

$$\hat{F} = (b - a)\frac{1}{N - 1}\sum_{i=0}^{N} f(X_i)$$

The random variable $X_i \in [a, b)$ can be constructed by warping a canonical random number uniformly distributed between zero and one, $\xi_i \in [0, 1)$: $X_i = a + \xi_i(b - a)$.

Markov chain - Monte Carlo technique.

Markov Chain Monte Carlo (MCMC) techniques are methods for sampling from probability distributions using Markov chains. MCMC methods are used in data modeling for Bayesian inference and numerical integration. Monte Carlo techniques are sampling methods.

Direct simulation: Let X be a random variable with distribution (x) ; then the expectation is given by:

$$E(X) = \sum_{x \in \mathcal{R}} x f(x)$$

which can be approximated by drawing n samples from $f(x)$ and then evaluating $E(X) \approx \frac{1}{n} \sum_{i=1}^{n} x_i$.

Thus, MCMC techniques aim to construct cleverly sampled chains which (after a burn in period) draw samples which are progressively more likely realizations of the distribution of interest; the target distribution.

11.8 Exercise

1. Define the concept MCMC techniques.

2. Obtain the value of pi using any simulation method.

11.9 Summary

Metropolis–Hastings algorithm: This method generates a Markov chain using a proposal density for new steps and a method for rejecting some of the proposed moves. It is actually a general framework which includes as special cases the very first and simpler MCMC (Metropolis algorithm) and many more recent alternatives listed below:

1. Gibbs sampling: This method requires all the conditional distributions of the target distribution to be sampled exactly. When drawing from the full-conditional distributions is not straightforward other samplers-within-Gibbs are used. Gibbs sampling is popular partly because it does not require any 'tuning'. Algorithm structure of the Gibbs sampling highly resembles that of the coordinate ascent variational inference in that both algorithms utilize the full-conditional distributions in the updating procedure

2. Metropolis-adjusted Langevin algorithm and other methods that rely on the gradient (and possibly second derivative) of the log target density to propose steps that are more likely to be in the direction of higher probability density
3. Pseudo-marginal Metropolis–Hastings: This method replaces the evaluation of the density of the target distribution with an unbiased estimate and is useful when the target density is not available analytically, e.g. latent variable models.

11.10 Further Readings

1. Berger, J.O. (1993) Statistical Decision Theory and Bayesian Analysis, Springer Verlag.
2. Gemerman, D and Lopes, H. F. (2006) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman Hall.
3. Leonard, T. and Hsu, J.S.J. (1999) Bayesian Methods, Cambridge University Press.
4. Robert, C.P. (1994). The Bayesian Choice: A Decision Theoretic Motivation, Springer.
5. Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods, Springer Verlag.
6. Lindley, D.V. (1965). Introduction to probability and statistical inference from Bayesian view point, Cambridge university press.