WILEY PUBLICATIONS IN STATISTICS

Statistical Decision Functions

Statistical Decision Functions

The Late ABRAHAM WALD
Professor of Mathematical Statistics
Columbia University

New York · John Wiley & Sons, Inc.
London

FIFTH PRINTING, MAY, 1964

PRINTED IN THE UNITED STATES OF AMERICA

Preface

This book presents the foundations of a recently developed general theory of statistical decision functions. It is mainly an outgrowth of several previous publications of the author on this subject and contains a considerable expansion and generalization of the ideas and results given in the earlier papers. A major advance beyond previous results is the treatment of the design of experimentation as a part of the general decision problem.

Until about ten years ago, the available statistical theories, except for a few scattered results, were restricted in two important respects: (1) experimentation was assumed to be carried out in a single stage; (2) the decision problems were restricted to problems of testing a hypothesis, and that of point and interval estimation. The general theory, as given in this book, is freed from both of these restrictions. It allows for multi-stage experimentation and includes the general multi-decision problem. A brief historical note on the developments leading up to the present stage of the theory is given in Section 1.7 of Chapter 1.

The first chapter is devoted to the formulation of the general decision problem and various basic concepts. It is shown that the decision problem may be interpreted as a zero sum two-person game in the sense of von Neumann's theory of games. The second chapter deals with a generalization of von Neumann's theory of zero sum two-person games, which is then used in Chapter 3 for the development of the theory of statistical decision functions. In Chapter 4 a number of additional results are given in the case of a sequence of identically and independently distributed chance variables. In Chapter 5 various special problems of interest are discussed, partly for the purpose of illustrating the general theory.

Throughout the book, general ideas and results are emphasized rather than specific methods or techniques. Some knowledge of probability, including probability distributions in the infinite dimensional space, is necessary for the understanding of the book. Because statistical concepts and ideas are developed in the book from the very beginning, a previous knowledge of statistics is not essential, although

still desirable. A knowledge of calculus and some familiarity with the elements of set, measure, and integration theories will suffice as a mathematical background for the reading of the book.

I am indebted to J. M. G. Fell, E. L. Lehmann, M. Loève, C. Stein, and J. Wolfowitz for reading the manuscript and for making valuable suggestions and remarks. The book was written under the sponsorship of the Office of Naval Research, and I wish to express my thanks for their generous support. Mrs. E. Bowker was most helpful in the preparation of the manuscript for publication, and I take this opportunity to thank her for her careful work.

A. W.

Columbia University
May, 1950

Contents

Chapter 1. THE GENERAL STATISTICAL DECISION PROBLEM: DEFINITIONS AND PRELIMINARY DISCUSSION

1.1 Formulation of the Statistical Decision Problem

1.1.1 The Stochastic Process Underlying the Statistical Decision Problem

Any statistical decision problem is formulated with reference to a stochastic process. By a stochastic process we mean a finite or infinite collection of chance variables having a joint probability distribution. We shall restrict ourselves to the case where the stochastic process consists of a countable collection of chance variables. Thus we shall assume that the stochastic process is given by a sequence $X = \{X_i\}$ ($i = 1, 2, \cdots$, ad inf.) of chance variables. For any sequence $x = \{x_i\}$ ($i = 1, 2, \cdots$, ad inf.) of real values, let $F(x)$ denote the probability that the inequalities $X_i < x_i$ hold simultaneously for all positive integral values i; i.e., $F(x)$ is the (cumulative) distribution function of X. Statistical decision problems with reference to the stochastic process X arise only when the distribution function $F(x)$ of X is not completely known. A characteristic feature of any statistical decision problem is the assumption that the unknown distribution $F(x)$ is merely known to be an element of a given class Ω of distribution functions. The class Ω is to be regarded as a datum of the decision problem; it will generally vary with the decision problem and the stochastic process under consideration. In most decision problems the class Ω will be a proper subset of the class of all possible distribution functions.

A frequent assumption in statistical problems is that the chance variables $X_1, X_2, \cdots$, etc., are independently and identically distributed. If this is all that is known about the distribution $F(x)$ of X, then the class Ω consists of all distribution functions $F(x)$ which can be written in the form $F(x) = \prod_{i=1}^{\infty} G(x_i)$, where $G(y)$ may be any univariate distribution function. In some problems merely the independence of the chance variables $X_1, X_2, \cdots$, etc., is postulated. The class Ω is then the class of all distribution functions $F(x)$ which can be written in the form $\prod_{i=1}^{\infty} G_i(x_i)$, the $G_i(x_i)$ being any univariate distribution functions. Much of the present-day statistical literature

deals with problems where Ω is a finite-parameter family of distribution functions. For example, if the chance variables $X_1, X_2, \cdots$, etc., are known to be independently distributed with the same normal distribution, but the mean and the standard deviation of the common normal distribution are unknown, then Ω will be a two-parameter family of distribution functions. Here is another simple example of a parametric class Ω: Suppose that it is known that for given values $x_1, \cdots, x_m$ of $X_1, \cdots, X_m$, respectively, the conditional distribution of X_{m+1} ($m = 1, 2, \cdots$, ad inf.) is normal with standard deviation σ and expected value $\alpha x_m + \beta$, where the values of the constants α, β, and σ are unknown. Suppose also that X_1 is known to be normally distributed with mean zero and standard deviation σ. Then Ω will be a three-parameter family of distribution functions.

1.1.2 Space of Possible Decisions at Termination of Experimentation

A statistical decision problem arises when we are faced with a set of alternative decisions, one of which must be made, and the degree of preference for the various possible decisions depends on the unknown distribution $F(x)$ of X.

As will be seen later, which of the possible decisions should be made will generally be determined only after some experimentation. By experimentation we mean making observations on some of the chance variables in the sequence $\{X_i\}$. Since the decisions under discussion here are made at the termination of experimentation, we shall refer to them as terminal decisions, as distinguished from decisions as to how to continue experimentation, which will be discussed in the next section. We shall use the symbol d^t to denote a terminal decision and the symbol D^t to denote the space of all possible terminal decisions d^t. In any decision problem there will be given a space D^t whose elements d^t represent the possible terminal decisions. The space D^t is to be regarded as a datum of the decision problem and will generally vary with the problem under consideration.

As an illustration, consider the following simple example. Suppose that a lot consisting of N units of a manufactured product is submitted for acceptance inspection. Suppose, in addition, that each unit is classified in one of two categories, defective or non-defective, and that the proportion p of defectives in the lot is unknown. We shall assign the value 1 to any defective unit, and 0 to any non-defective unit. The two possible terminal decisions under consideration here are acceptance or rejection of the lot. Obviously the degree of preference for acceptance or rejection of the lot will depend on the proportion p

of defectives in the lot. In general it will be possible to specify a value p_0 such that acceptance is preferred when $p < p_0$ and rejection is preferred when $p \geqq p_0$. A decision problem arises if complete inspection of the lot is too costly and we have to decide on acceptance or rejection on the basis of a limited random sample drawn from the lot. For this problem the space D^t consists of the two elements ${d_1}^t$ and ${d_2}^t$, where ${d_1}^t$ denotes the decision to accept the lot, and ${d_2}^t$ the decision to reject the lot. The stochastic process underlying this decision problem consists of the finite sequence $\{X_i\}$ $(i = 1, \cdots, N)$ of chance variables corresponding to successive random drawings from the lot without replacement (X_i corresponds to the ith drawing). The joint distribution of $X_1, \cdots, X_N$ is determined as follows: Each X_i can take only the values 0 and 1. The probability that $X_1 = 1$ is equal to p. The conditional probability that $X_m = 1$, given that $X_1 = x_1, \cdots, X_{m-1} = x_{m-1}$, is equal to $\dfrac{pN - \sum_1^{m-1} x_i}{N - m + 1}$. Thus Ω in this problem is a one-parameter family of distribution functions, the only unknown parameter being p. Experimentation consisting of the inspection of m units drawn from the lot means making observations on the first m chance variables $X_1, \cdots, X_m$.

In general it will be possible to associate each element d^t of the space D^t with some subset ω of Ω such that the decision d^t can be interpreted as the decision to accept the hypothesis that the true distribution $F(x)$ of X is an element of ω. For instance, in the example discussed above the decision ${d_1}^t$ can be interpreted as the decision to accept the hypothesis that $p < p_0$, and ${d_2}^t$ as the decision to accept the hypothesis that $p \geqq p_0$.

Suppose that for any element F of Ω and for any two elements ${d_1}^t$ and ${d_2}^t$ of D^t one (and only one) of the following three statements is true: (1) ${d_1}^t$ is preferred to ${d_2}^t$ when F is true; (2) ${d_2}^t$ is preferred to ${d_1}^t$ when F is true; (3) neither of the two decisions ${d_1}^t$ or ${d_2}^t$ is preferred when F is true.

An element d^t of D^t may be called optimal relative to an element F of Ω if there is no element of D^t that is preferred to d^t when F is true. With each element d^t we associate the set ω_{d^t} of all elements F of Ω relative to which d^t is optimal. In general it will be possible to interpret d^t as the decision to accept the hypothesis that the true distribution F is an element of ω_{d^t}, provided that ω_{d^t} is not empty. Although in most decision problems the set ω_{d^t} will not be empty for any d^t, there are problems, not without interest, where ω_{d^t} is empty for some d^t and the above-mentioned interpretation of d^t becomes meaningless.

In most of the problems treated so far in statistical literature, each element d^t of D^t is defined from the outset as the decision to accept the hypothesis that F is an element of a certain subset ω of Ω. While this is undoubtedly the most important case to be considered, we do not wish to restrict the generality of our investigations by imposing such a condition on the nature of the elements d^t.

1.1.3 Space of Possible Decisions as to How to Continue Experimentation at Any Given Stage

As mentioned in Section 1.1.2, by experimentation we mean making observations on some of the chance variables in the sequence $\{X_i\}$ ($i = 1, 2, \cdots$, ad inf.). It will be assumed that at most one observation is made on each chance variable X_i. There is no loss of generality in making this assumption. Suppose, for example, that the experimenter makes $r (r > 1)$ independent observations on X_i, say $X_{i1}, \cdots, X_{ir}$; then X_i can be replaced by a finite set of independently and identically distributed chance variables $X_{i1}, \cdots, X_{ir}$, and x_{ij} can be regarded as a single observation on X_{ij}.[1]

We shall permit experimentation to be carried out in several stages. The first stage consists of the selection of a certain finite set of chance variables from the sequence $\{X_i\}$ and observation of their values. After the first stage has been completed, the second stage is carried out by selecting a finite set from the remaining chance variables in the sequence $\{X_i\}$ and observing their values, and so on. If experimentation is terminated after the kth stage, we shall say that the experiment has been carried out in k stages. Experimentation in several stages is frequently preferable to experimentation in a single stage, since in the former type of experimentation the selection of the chance variables to be observed in the next stage may be made dependent on the observed values obtained in all the preceding stages.

Before the start of experimentation, the experimenter is confronted with the following question of choice for the first stage of experimentation: Which finite group of elements of the sequence $\{X_i\}$ should he observe? Thus any decision concerning the first stage of experimentation can be represented by a finite set of positive integers $i_1, \cdots, i_k$ which are pairwise different. The set $\{i_1, \cdots, i_k\}$ represents the decision to make an observation on each of the chance variables $X_{i_1}, \cdots,$

[1] More generally, the sequence $\{X_i\}$ can be replaced by the double sequence $\{Y_{ij}\}$ ($i, j = 1, 2, \cdots$, ad inf.) of chance variables, where the distribution of $Y_i = \{Y_{ij}\}$ ($j = 1, 2, \cdots$, ad inf.) is identical with that of $X = \{X_j\}$ ($j = 1, 2, \cdots$, ad inf.) and $Y_1, Y_2, \cdots$, etc., are independent. The double sequence $\{Y_{ij}\}$ can be arranged in a single sequence $Z = \{Z_i\}$, and we may regard Z as the stochastic process underlying the decision problem.

X_{i_k}. Consider now a later stage of experimentation when observations have already been made on $X_{j_1}, \cdots, X_{j_r}$ but on no other chance variables. Then any possible decision to continue experimentation one stage further can be represented by a finite subset of $I - \{j_1, \cdots, j_r\}$, where I denotes the set of all positive integers. If $h_1, \cdots, h_m$ are elements of $I - \{j_1, \cdots, j_r\}$, then the set $\{h_1, \cdots, h_m\}$ represents the decision to observe $X_{h_1}, \cdots, X_{h_m}$.

At this point, one may raise the question why a single stage of experimentation should consist of more than one observation. On first thought, it may seem more reasonable to select merely one chance variable for observation at a time and to make further selections of chance variables dependent on the observed value of that chance variable. There are situations, however, where such a procedure would be rather costly and impractical. For example, if making an observation requires a considerable amount of time, as it frequently does in agricultural experimentation, the selection for observation of merely one chance variable at a time may make the time needed for the completion of the experiment so long as to make its value almost worthless. There may also be other reasons why the selection of more than one chance variable at a time may be desirable.

Let D^e be the space of all possible decisions as to the first stage of experimentation; i.e., D^e is the space of all finite (non-empty) subsets of the set I of all positive integers. Thus any element d^e of D^e is simply a finite (non-empty) subset of I. After observations have been made on $X_{i_1}, \cdots, X_{i_k}$, but on no other chance variables, the space $D^e{}_{i_1 \cdots i_k}$ of all possible decisions on the next stage of experimentation, if experimentation is to be continued at all, consists of all elements d^e of D^e which are subsets of the set $I - \{i_1, \cdots, i_k\}$.

As an illustration, consider the following example: Suppose that the elements of $\{X_i\}$ $(i = 1, 2, \cdots)$ are independent, those with odd subscripts are normally distributed with mean θ_1 and variance 1, and those with even subscripts are normally distributed with mean θ_2 and variance 9. The values of the parameters θ_1 and θ_2 are assumed to be unknown. Let the decision space D^t consist of the two elements $d_1{}^t$ and $d_2{}^t$, where $d_1{}^t$ denotes the decision to accept the hypothesis H that $\theta_1 < \theta_2$, and $d_2{}^t$ the decision to reject the hypothesis H. Since the X's with odd subscripts, as well as those with even subscripts, are identically distributed, the question as to how experimentation should be carried out reduces to this: How many X's with odd subscripts, and how many with even subscripts, should be observed in the first stage of the experiment? After the first stage has been completed, the question is again how many X's with odd, and how many with even,

subscripts should be observed in the second stage of the experiment, and so on. Since the standard deviation of an X with even index is 3 and that of an X with odd index is 1, it is intuitively clear that it will be advantageous to observe more X's with even subscript than X's with odd subscript, provided that the cost of observing the value of X_i is independent of i.

1.1.4 Decision Functions

We are now in a position to define the notion of a decision function. First we shall give the definition of a special type of decision function, the so-called non-randomized decision function. Let D be the set-theoretical sum of D^t and D^e; i.e., D consists of all elements d^t of D^t and all elements d^e of D^e. Furthermore, for any subset $\{i_1, \cdots, i_k\}$ of the set I of all positive integers, let $D_{i_1 \cdots i_k}$ be the set-theoretical sum of D^t and $D^e{}_{i_1 \cdots i_k}$. A function $d(x; s_1, \cdots, s_k)$ is said to be a non-randomized decision function if: (1) it is a single-valued function defined for all positive integral values k, for any sample point x, and for any finite disjoint sets $s_1, \cdots, s_k$ of positive integers; (2) the value of $d(x; s_1, \cdots, s_k)$ is independent of the coordinates x_i of x for which the integer i is not contained in any of the sets $s_1, \cdots, s_k$; (3) it is a constant when $k = 0$ [we shall denote this constant by $d(0)$]; (4) for $k \geqq 1$, the value of the function $d(x; s_1, \cdots, s_k)$ may be any element of $D_{i_1 \cdots i_r}$, where the set $\{i_1, \cdots, i_r\}$ is the set-theoretical sum of $s_1, \cdots, s_k$; (5) for $k = 0$, the value of $d(x; s_1, \cdots, s_k)$, i.e., the value $d(0)$, may be any element of D.

Such a decision function can be used to determine uniquely a rule for carrying out the experimentation and for selecting a terminal decision d^t. This can be done as follows: If $d(0)$ is an element d^t of D^t, no experimentation is made and the terminal decision $d(0)$ is chosen. If $d(0)$ is an element $d^e = s_1 = (i_1, \cdots, i_r)$ of D^e, then observations are made on the chance variables $X_{i_1}, \cdots, X_{i_r}$ and the value of $d(x, s_1)$ is computed. If $d(x; s_1)$ is an element d^t of D^t, experimentation is stopped and the terminal decision $d(x; s_1)$ is made. If $d(x; s_1)$ is an element $d^e = s_2 = (j_1, \cdots, j_u)$, then observations are made on $X_{j_1}, \cdots, X_{j_u}$ and the value of $d(x; s_1, s_2)$ is computed. If $d(x; s_1, s_2)$ is an element d^t of D^t, experimentation is stopped with the terminal decision $d(x; s_1, s_2)$. If $d(x; s_1, s_2)$ is an element of D^e, observations are made on the corresponding set of chance variables, and so on.

Let C_D be a certain Borel field [2] of subsets of the space D which

[2] A class C of subsets of a space A is said to be a Borel field if (i) the empty set belongs to C; (ii) if a subset α of A belongs to C, then the complement of α also belongs to C; (iii) the sum of a sequence $\{\alpha_i\}$ ($i = 1, 2, \cdots$, ad inf.) of subsets of A belongs to C if α_i belongs to C for each i.

contains all denumerable subsets of D as elements. By a probability measure δ on the space D we shall mean a probability measure defined for all elements of the Borel field C_D. Let Δ be the space of all probability measures δ. For any subset $\{i_1, \cdots, i_k\}$ of I, let $\Delta_{i_1 \cdots i_k}$ be the class of all probability measures δ for which $\delta(D_{i_1 \cdots i_k}) = 1$.

A function $\delta(x; s_1, \cdots, s_k)$ whose values are elements of Δ is said to be a randomized decision function if: (1) it is a single-valued function defined for any positive integer k, for any finite disjoint sets $s_1, \cdots, s_k$ of positive integers, and for any sample point x; (2) it is a constant $\delta(0)$ when $k = 0$; (3) $\delta(x; s_1, \cdots, s_k)$ is an element of $\Delta_{i_1 \cdots i_r}$ if $r \geqq 1$, and $\delta(0)$ is an element of Δ where $\{i_1, \cdots, i_r\}$ is the set-theoretical sum of $s_1, \cdots, s_k$; (4) the value of $\delta(x; s_1, \cdots, s_k)$ is independent of the coordinates x_i of x for which the integer i is not contained in any of the sets $s_1, \cdots, s_k$.

Clearly, a randomized decision function is equivalent to a non-randomized decision function if for any values k, $s_1, \cdots, s_k$, and x the probability measure $\delta(x; s_1, \cdots, s_k)$ assigns the probability 1 to a single element of D. Thus a non-randomized decision function may be regarded as a special case of a randomized decision function.[3] A randomized decision function $\delta(x; s_1, \cdots, s_k)$ can also be used to determine uniquely a procedure for making the experimentation and selecting a terminal decision. First an element d of D is selected with the help of a chance mechanism constructed so that the probability distribution of the selected element d is equal to $\delta(0)$. If the element d so selected is a terminal decision d^t, no experimentation is made and the terminal decision d^t is adopted. If the element d so selected is an element $d^e = s_1 = (i_1, \cdots, i_r)$ of D^e, then observations are made on $X_{i_1}, \cdots, X_{i_r}$ and the value of $\delta(x; s_1)$ is computed. Then the probability distribution $\delta(x; s_1)$ is used to select an element d of D. If the element d so selected is contained in D^t, experimentation is stopped with the corresponding terminal decision. If it is an element of D^e, observations are made on the corresponding set of chance variables, and so on. The procedure is the same as in the non-randomized case, except that at each stage, instead of choosing a particular element d, the experimenter chooses a probability measure δ on D and then the element d is selected with the help of a chance mechanism that produces the desired probability distribution δ.

It would seem reasonable to assume that $\delta(x; s_1, \cdots, s_k) = \delta(x; s'_1, \cdots, s'_r)$ if the set-theoretical sum of $s_1, \cdots, s_k$ is equal to that of $s'_1, \cdots, s'_r$. We shall, however, not make this restriction on δ for reasons that will be apparent in Chapter 3.

[3] We may identify "selection with probability 1" with "selection with certainty."

In what follows the term "decision function" will be used for randomized as well as for non-randomized decision functions, since the latter are a special case of the former. For any subset D^* of D, we shall denote the probability that $d \epsilon D^*$ by $\delta(D^* \mid x; s_1, \cdots, s_k)$ when $\delta(x; s_1, \cdots, s_k)$ $(k > 0)$ is the probability measure on D, and by $\delta(D^* \mid 0)$ when $\delta(0)$ is the probability measure on D.

An important special case arises when the decision function $\delta(x; s_1, \cdots, s_k)$ used is such that it is certain that experimentation is carried out exactly in one stage. This will be the situation when $\delta(D^{\mathfrak{C}} \mid 0) = 1$ and $\delta(D^t \mid x; s_1) = 1$ for any x and s_1. We can characterize this case also by saying that we decide in advance (before experimentation starts) which chance variables in the sequence $\{X_i\}$ should be observed during the total course of experimentation. This is the classical non-sequential case. A decision function $\delta(x; s_1, \cdots, s_k)$ will be said to be sequential if it is such that, if adopted by the experimenter, the probability is positive that the experiment will be carried out in more than one stage.

1.1.5 Losses Due to Possible Wrong Terminal Decisions and Cost of Experimentation

The experimenter is confronted with the problem of choosing a particular decision function $\delta(x; s_1, \cdots, s_k)$ for carrying out the experimentation and making a terminal decision. But, to be able to judge the relative merit of any given decision function, it is necessary that something be stated about (1) the relative degree of preference given to the various elements d^t of D^t when the true distribution F of X is known, and (2) the cost of experimentation.

The degree of preference given to the various elements d^t of D^t when F is known can be expressed by a non-negative function $W(F, d^t)$, called weight function, which is defined for all elements F of Ω and all elements d^t of D^t. For any pair (F, d^t), the value of $W(F, d^t)$ expresses the loss suffered by making the terminal decision d^t when F is the true distribution of X. We shall say that d^t is a correct terminal decision when F is true, if $W(F, d^t)$ is zero. If $W(F, d^t) > 0$, we shall say that d^t is a wrong terminal decision when F is true. A terminal decision ${d_1}^t$ is said to be preferable to another terminal decision ${d_2}^t$ when F is true, if $W(F, {d_1}^t) < W(F, {d_2}^t)$.

The weight function $W(F, d^t)$ is to be regarded as a datum of the problem. In some problems, however, it may be difficult to set up a numerical weight function $W(F, d^t)$, especially when d^t means the acceptance or rejection of a certain scientific hypothesis. Even in those cases where there is no difficulty in principle in assigning a numer-

ical value to $W(F, d^t)$ for any F and d^t, the resulting weight function may be rather complicated and it would be desirable to replace it by some simplified function. We shall say that a weight function $W(F, d^t)$ is simple if it can take only the values 0 and 1. In many statistical problems it will be sufficient for practical purposes to consider only simple weight functions. In problems where there is difficulty in assigning definite numerical values to losses due to terminal decisions, there will generally be no such difficulty in constructing a simple weight function, since the latter merely requires that for any given F the elements d^t of D^t be classified in two categories only, wrong and correct decisions. If a numerical weight function $W(F, d^t)$ can be constructed, but we wish to replace it by a simple weight function $W^*(F, d^t)$ for reasons of simplicity, we may proceed as follows: For any F let $c(F)$ be some properly chosen positive value. We then put $W^*(F, d^t) = 0$ when $W(F, d^t) < c(F)$, and $W^*(F, d^t) = 1$ when $W(F, d^t) \geqq c(F)$.

As an illustration, consider the example discussed in Section 1.1.2. In that example the space D^t consists of two elements ${d_1}^t$ and ${d_2}^t$, where ${d_1}^t$ denotes the decision to accept the lot and ${d_2}^t$ the decision to reject the lot. Since the proportion p of defectives in the lot is the only parameter on which F depends, we may replace F by p in the weight function $W(F, d^t)$. Consider the following weight function:

$$W(p, {d_1}^t) = 0 \quad \text{for } p \leqq p_0, \qquad = c_1(p - p_0) \quad \text{for } p > p_0$$

$$W(p, {d_2}^t) = c_2(p_0 - p) \quad \text{for } p \leqq p_0, \qquad = 0 \quad \text{for } p > p_0$$

It will generally be possible to choose the constants p_0, c_1, and c_2 so that the resulting weight function will express the preference scale sufficiently well for practical purposes. If we want to replace the above weight function by a simple weight function $W^*(p, d^t)$, we choose two values p_1 and $p_2 (p_1 < p_0 < p_2)$ and put

$$W^*(p, {d_1}^t) = 0 \quad \text{when } p \leqq p_2, \qquad = 1 \quad \text{when } p > p_2$$

$$W^*(p, {d_2}^t) = 1 \quad \text{when } p \leqq p_1, \qquad = 0 \quad \text{when } p > p_1$$

Such a simple weight function, if p_1 and p_2 are chosen properly, will frequently be satisfactory for practical purposes.

The cost of experimentation may depend on the chance variables selected for observation, on the actual observed values obtained, and also on the stages in which the experiment has been carried out. Thus we shall denote the cost by $c(x; s_1, \cdots, s_k)$ when (1) the experiment was carried out in k stages; (2) the ith stage consisted of the observations on the chance variables X_j for all j that are elements of s_i; (3) x is

the observed sample point. Of course, the cost $c(x; s_1, \cdots, s_k)$ does not depend on the coordinates x_i of x for which i is not contained in any of the sets $s_1, \cdots, s_k$.

A special case of interest is that where the cost of experimentation depends only on the number of observations made and is proportional to it. A cost function of this sort will be called a simple cost function. In many problems it will be possible to approximate the cost function by a simple one.

1.1.6 Statement of the Decision Problem

We are now in a position to give a formulation of the general decision problem. It may be stated as follows:

Given (1) the stochastic process $\{X_t\}$, (2) the class Ω of distributions which is known to contain the true distribution F of X as an element, (3) the space D^t of possible terminal decisions, (4) the weight function $W(F, d^t)$ defined for all elements F of Ω and all elements d^t of D^t, and (5) the cost function $c(x; s_1, \cdots, s_k)$ of experimentation, the problem is to choose a decision function $\delta(x; s_1, \cdots, s_k)$ to be adopted for carrying out the experiment and for making a terminal decision.

The adoption of a particular decision function by the experimenter may be termed "inductive behavior," since it determines uniquely the procedure for carrying out the experiment and for making a terminal decision. Thus the above decision problem may be called the problem of inductive behavior.[4]

In attempting to solve the above decision problem, the first essential step is to set up some principles which will lead to a complete, or at least a partial, ordering of all possible decision functions with respect to their suitability for purposes of inductive behavior. This will be done in Sections 1.2 and 1.3 by introducing the notions of uniformly better decision functions and admissible decision functions.

1.2 Consequences of the Adoption of a Particular Decision Function

1.2.1 The Risk Function

First we shall introduce some notation that will prove to be convenient. Let $s_1, \cdots, s_r$ be r disjoint subsets of the set I of all positive integers, and let s denote the sequence $\{s_1, \cdots, s_r\}$ of the sets $s_1, \cdots, s_r$. For any sequence $x = \{x_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of real numbers we shall use $\delta(x; s)$ as an alternative notation for $\delta(x; s_1, \cdots, s_r)$ and

[4] The term "inductive behavior" was introduced by Neyman [38]. (The number in brackets refers to an item in the Bibliography at the end of the book.)

$c(x; s)$ as an alternative notation for $c(x; s_1, \cdots, s_r)$; i.e., we put

$$\delta(x; s_1, \cdots, s_r) = \delta(x; s) \tag{1.1}$$

$$c(x; s_1, \cdots, s_r) = c(x; s) \tag{1.2}$$

In accordance with the notation in (1.1), for any subset D^* of D, we shall use the symbols $\delta(D^* \mid x; s_1, \cdots, s_r)$ and $\delta(D^* \mid x; s)$ synonymously to denote the probability that the decision d made will be contained in D^* when the decision function δ is used, x is the observed sample, and r stages of the experiment have been carried out in accordance with $s_1, \cdots, s_r$, respectively.

We shall occasionally use the same symbol, d^e, to denote a given element of D^e, as well as to denote the set of positive integers by which this element of D^e is represented, provided that this can be done without any danger of confusion.

If the decision function $\delta(y; s)$ is adopted and if $x = \{x_i\}$ is the observed sample point, i.e., x_i is the observed value of X_i, then the probability that the experiment will be carried out in k stages, the first stage in accordance with $d_1{}^e$, the second in accordance with $d_2{}^e, \cdots$, the kth stage in accordance with $d_k{}^e$, and that the terminal decision will be an element of the subset $\bar{D}^t$ of D^t is given by

$$p(d_1{}^e, d_2{}^e, \cdots, d_k{}^e, \bar{D}^t \mid x, \delta) = \delta(d_1{}^e \mid 0)\delta(d_2{}^e \mid x; d_1{}^e) \tag{1.3}$$

$$\delta(d_3{}^e \mid x; d_1{}^e, d_2{}^e) \cdots \delta(d_k{}^e \mid x; d_1{}^e, \cdots, d^e{}_{k-1})\delta(\bar{D}^t \mid x; d_1{}^e, \cdots, d_k{}^e)$$

For $k = 0$, the right-hand member of (1.3) reduces to $\delta(\bar{D}^t \mid 0)$.

The probability of the same event when merely the adopted decision function $\delta(y; s)$ but not the observed sample point x is given, and when F is the true distribution, is equal to [5]

$$q(d_1{}^e, \cdots, d_k{}^e, \bar{D}^t \mid F, \delta) = \int_M p(d_1{}^e, \cdots, d_k{}^e, \bar{D}^t \mid x, \delta)\, dF(x) \tag{1.4}$$

where M denotes the whole sample space; i.e., M is the totality of all sequences x. Thus the probability that the terminal decision will be an element of $\bar{D}^t$ when δ is used and F is true is given by

$$P(\bar{D}^t \mid F, \delta) = \sum_{k=0}^{\infty} \sum_{d_1{}^e, \cdots, d_k{}^e} q(d_1{}^e, \cdots, d_k{}^e, \bar{D}^t \mid F, \delta) \tag{1.5}$$

[5] The integral in (1.4) has a meaning only if $p(d_1{}^e, \cdots, d_k{}^e, \bar{D}^t \mid x, \delta)$ is a measurable function of x. The precise measurability conditions which will insure the existence of this integral, as well as that of the integrals which will appear in subsequent formulas of this chapter, will be stated in Chapter 3.

Hence the expected value of the loss $W(F, d^t)$ when δ is used and F is true is equal to

$$r_1(F, \delta) = \int_{D^t} W(F, d^t)\, dP(\bar{D}^t \mid F, \delta) \tag{1.6}$$

The expected value of the cost of experimentation when F is true and the decision function $\delta(y; s)$ is used is given by [6]

$$r_2(F, \delta) = \sum_{k=1}^{\infty} \sum_{d_1^e, \cdots, d_k^e} \int_M c(x; d_1^e, d_2^e, \cdots, d_k^e)\, p(d_1^e, \cdots, d_k^e, D^t \mid x, \delta)\, dF(x) \tag{1.7}$$

The sum of the expected value of $W(F, d^t)$ and the expected cost of experimentation is called the risk; i.e., the risk is given by

$$r(F, \delta) = r_1(F, \delta) + r_2(F, \delta) \tag{1.8}$$

The risk $r(F, \delta)$ will be called simple risk if the underlying weight and cost functions are simple.

It seems reasonable to judge the merit of any given decision function δ_0 for purposes of inductive behavior entirely on the basis of the risk function $r(F, \delta_0)$ associated with it. This already permits a partial ordering of the decision functions as to their suitability for purposes of inductive behavior. Clearly, if the merit of any decision function is judged entirely on the basis of its risk function, the decision function δ_1 will be preferred to the decision function δ_2 if the following inequalities hold:

$$r(F, \delta_1) \leqq r(F, \delta_2) \tag{1.9}$$

for all F in Ω, and

$$r(F, \delta_1) < r(F, \delta_2) \tag{1.10}$$

for at least one element F of Ω.

If the above inequalities hold, we shall also say that δ_1 is uniformly better than δ_2.

As an illustration, we shall discuss briefly the following simple example. Let $X_1, X_2, \cdots$, ad inf., be independently and normally distributed chance variables with variance 1 and common mean θ, the value of which is unknown. Suppose that the space D^t of terminal decisions contains only two elements $d_1{}^t$ and $d_2{}^t$, where $d_1{}^t$ is the decision to accept the hypothesis H that $\theta < 0$, and $d_2{}^t$ is the decision to

[6] Formula (1.7) is valid if the probability is 1 that the experiment will be carried out in a finite number of stages. Otherwise, $r_2(F, \delta) = \infty$, since it will be assumed in Chapter 3 that the cost of making infinitely many observations is ∞. The cost of taking no observations is assumed to be zero.

reject the hypothesis H. In this case, Ω is a one-parameter family of distributions, since each element F of Ω is determined by a particular value of θ. We shall assume that the cost of experimentation is proportional to the number of observations made and that the weight $W(\theta, d^t)$ is given as follows: $W(\theta, d_1{}^t) = 0$ when $\theta \leqq \rho$ and $= 1$ when $\theta > \rho$; $W(\theta, d_2{}^t) = 1$ when $\theta < -\rho$ and $= 0$ when $\theta \geqq -\rho$, where ρ is a given positive value. Thus we have a simple weight function and a simple cost function. Consider now the particular decision functions δ_1 and δ_2 defined as follows: $\delta_1(0)$ assigns the probability 1 to the element $d^e = (1, 2, \cdots, 9)$, and $\delta_1[x_1, \cdots, x_9; (1, 2, \cdots, 9)]$ assigns the probability 1 to $d_1{}^t$ or $d_2{}^t$ according to whether $\bar{x} = (x_1 + \cdots + x_9)/9 \leqq 0$ or >0. The probability measure $\delta_2(0)$ assigns the probability 1 to $d^e = (1, 2, \cdots, 9)$, and $\delta_2[x_1, \cdots, x_9; (1, 2, \cdots, 9)]$ assigns the probability 1 to $d_1{}^t$ or $d_2{}^t$ according to whether the median $\tilde{x}$ of $(x_1, \cdots, x_9)$ is $\leqq 0$ or >0. Thus, if δ_1 is adopted, the experimenter makes one observation x_i on X_i for $i = 1, 2, \cdots, 9$, and accepts H if $\bar{x} \leqq 0$ or rejects H if $\bar{x} > 0$. If δ_2 is adopted, again one observation x_i is made on X_i for each $i \leqq 9$, and then H is accepted if $\tilde{x} \leqq 0$ and H is rejected if $\tilde{x} > 0$. We shall now compute the risk functions associated with δ_1 and δ_2. Let $G(y)$ be the Gaussian distribution function; i.e.,

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2}\, dt \tag{1.11}$$

Clearly the probability that the terminal decision $d_1{}^t$ will be made when θ is true and δ_1 is used is given by

$$P(d_1{}^t \mid \theta, \delta_1) = G(-3\theta) \tag{1.12}$$

and the probability of the same event when θ is true and δ_2 is used is given by

$$P(d_1{}^t \mid \theta, \delta_2) = \sum_{j=5}^{9} \binom{9}{j} [G(-\theta)]^j [1 - G(-\theta)]^{9-j} \tag{1.13}$$

The risk associated with δ_i $(i = 1, 2)$ is equal to

$$\begin{aligned} r(\theta, \delta_i) &= 9c + P(d_2{}^t \mid \theta, \delta_i) && \text{when } \theta < -\rho \\ &= 9c && \text{when } -\rho \leqq \theta \leqq \rho \\ &= 9c + P(d_1{}^t \mid \theta, \delta_i) && \text{when } \theta > \rho \end{aligned} \tag{1.14}$$

where c is the cost of a single observation and $P(d_2{}^t \mid \theta, \delta_i) = 1 - P(d_1{}^t \mid \theta, \delta_i)$. Clearly $r(\theta, \delta_1) = r(\theta, \delta_2)$ when $|\theta| \leqq \rho$. One

can verify that $r(\theta, \delta_1) < r(\theta, \delta_2)$ when $|\theta| > \rho$. Thus δ_1 is uniformly better. The decision functions δ_1 and δ_2 are of the classical type, since according to both decision functions experimentation is carried out in one stage. Since the cost function assumed in this example does not depend on the number of stages in which the experiment is carried out, a reduction of the risk is possible if one uses decision functions for which the probability is positive that the experiment will be carried out in more than one stage. Consider, for example, the decision function δ_3 defined as follows: $\delta_3(0)$ assigns the probability 1 to $d^e = (1, 2, 3, 4)$. $\delta_3[x_1, x_2, x_3, x_4; (1, 2, 3, 4)]$ assigns the probability 1 to $d^e = (5, 6, 7, 8, 9)$ if $-a < (x_1 + \cdots + x_4)/4 < a$, the probability 1 to $d_1{}^t$ if $(x_1 + \cdots + x_4)/4 \leqq -a$, and the probability 1 to $d_2{}^t$ if $(x_1 + \cdots + x_4)/4 \geqq a$, where a is a given positive number. Furthermore we put $\delta_3[x_1, \cdots, x_9; (1, 2, 3, 4), (5, 6, 7, 8, 9)] = \delta_1[x_1, \cdots, x_9; (1, \cdots, 9)]$. Clearly the expected cost of experimentation associated with δ_3 will be smaller than that associated with δ_1, the reduction being considerable when $|\theta| \geqq a$. On the other hand, if a is sufficiently large, the expected value of the loss $W(\theta, d^t)$ when δ_3 is used will for all practical purposes coincide with that corresponding to δ_1.

1.2.2 The Performance Characteristic

The probability $P(\bar{D}^t \mid F, \delta)$ that the terminal decision will be an element of a given subset $\bar{D}^t$ of D^t when F is true and the decision function δ is adopted becomes a function of the two variables $\bar{D}^t$ and F if δ is specified. For any particular δ, say δ_0, we shall call the function $P(\bar{D}^t \mid F, \delta_0)$ the performance characteristic of δ_0 regarding terminal decisions.

Let $q(d_1{}^e, \cdots, d_k{}^e \mid F, \delta)$ be equal to $q(d_1{}^e, \cdots, d_k{}^e, D^t \mid F, \delta)$, where $q(d_1{}^e, \cdots, d_k{}^e, \bar{D}^t \mid F, \delta)$ is the function defined in (1.4). Thus $q(d_1{}^e, \cdots, d_k{}^e \mid F, \delta)$ is the probability, when F is true and δ is used, that the experiment will be carried out in k stages, the first stage in accordance with $d_1{}^e$, $\cdots$, the kth stage in accordance with $d_k{}^e$. For any given δ, say δ_0, the function $q(d_1{}^e, \cdots, d_k{}^e \mid F, \delta_0)$ will depend only on k, $d_1{}^e, \cdots, d_k{}^e$, and F. This function will be called the performance characteristic of δ_0 regarding experimentation.

The performance characteristic regarding terminal decisions determines uniquely the expected value of $W(F, d^t)$ for any given weight function $W(F, d^t)$ [see formula (1.6)]. The performance characteristic regarding experimentation determines uniquely the expected cost of experimentation for any given cost function, provided that the cost of experimentation $c(x; d_1{}^e, \cdots, d_k{}^e)$ does not depend on x; i.e.,

$c(x; d_1^e, \cdots, d_k^e) = c(d_1^e, \cdots, d_k^e)$, which will be the case in most problems arising in applications.

1.3 Admissible Decision Functions and Complete Classes of Decision Functions

A decision function δ will be said to be admissible if there exists no other decision function δ^* which is uniformly better than δ, i.e., if there exists no decision function δ^* satisfying the following two conditions:

$$r(F, \delta^*) \leqq r(F, \delta) \tag{1.15}$$

for all F in Ω, and

$$r(F, \delta^*) < r(F, \delta) \tag{1.16}$$

for at least one element F of Ω.

A class C of decision functions δ will be said to be complete if for any δ not in C we can find an element δ^* in C such that δ^* is uniformly better than δ.

A complete class C will be said to be a minimal complete class if no proper subclass of C is complete. If a minimal complete class exists, it must be equal to the class C_0 of all admissible decision functions. This can be seen as follows: Let C_1 be a minimal complete class. Clearly C_0 must be a subset of C_1. Suppose that there exists an element δ' of C_1 that is not an element of C_0. Then there exists a decision function δ'' which is uniformly better than δ'. Since C_1 is a minimal complete class, δ'' cannot be an element of C_1. But then there exists an element δ''' in C_1 that is uniformly better than δ'' and, therefore, also uniformly better than δ', which is not possible since C_1 is a minimal complete class. Thus $C_1 = C_0$.

If the class C_0 of all admissible decision functions is complete, it is evidently a minimal complete class. Since no minimal complete class exists that is different from C_0, a necessary and sufficient condition for the existence of a minimal complete class is that C_0 be complete.

As will be seen in Chapter 3, the class C_0 will be complete under very general conditions. Exceptional cases may arise, for example, when the space D^t is not complete in the following sense: there exists a sequence $\{d_i^t\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ such that $\lim_{i=\infty} W(F, d_i^t) = W(F)$ but no element d^t exists such that $W(F, d^t) = W(F)$.

The notions of admissibility and complete classes are of basic importance in the theory of decision functions. They will be studied in Chapter 3.

1.4 Bayes and Minimax Solutions of the Decision Problem

1.4.1 Decision Functions which Minimize Some Average Risk (Bayes Solutions)

Let C_Ω be a Borel field of subsets of Ω which contains all denumerable subsets of Ω as elements. By a probability measure ξ on Ω we shall mean a probability measure defined for all elements of C_Ω. The question of how to choose C_Ω will be discussed in Chapter 3. A probability measure ξ in Ω will also be called an *a priori distribution in* Ω.

If an a priori distribution ξ in Ω exists and is known to the experimenter, a decision function for which the average risk (averaged with the a priori distribution ξ), i.e., the expression

$$\int_\Omega r(F, \delta)\, d\xi = r^*(\xi, \delta) \tag{1.17}$$

takes its minimum value may be regarded as an optimum solution. A decision function δ_0 which minimizes $r^*(\xi, \delta)$, i.e., for which

$$r^*(\xi, \delta_0) \leqq r^*(\xi, \delta) \tag{1.18}$$

for all δ is called a Bayes solution relative to the a priori distribution ξ.

Let ξ_F be the particular a priori distribution which assigns the probability one to the element F of Ω. Then obviously we have

$$r(F, \delta) = r^*(\xi_F, \delta) \tag{1.19}$$

Thus we can interpret the value of $r(F, \delta)$ as the value of $r^*(\xi_F, \delta)$. In what follows we shall write $r(\xi, \delta)$ for $r^*(\xi, \delta)$, and $r(F, \delta)$ will be used synonymously with $r(\xi_F, \delta)$. This can be done without any danger of confusion.

In many statistical problems the existence of an a priori distribution cannot be postulated, and, in those cases where the existence of an a priori distribution can be assumed, it is usually unknown to the experimenter and therefore the Bayes solution cannot be determined. The main reason for discussing Bayes solutions here is that they enter into some of the basic results in Chapter 3. It will be shown there that under certain rather weak conditions the class of all Bayes solutions corresponding to all possible a priori distributions is a complete class.

Let $\{\xi_i\}$ ($i = 1, 2, \cdots,$ ad inf.) be a sequence of a priori distributions and δ_0 a decision function. We shall say that δ_0 is a Bayes solution relative to the sequence $\{\xi_i\}$ if

$$\lim_{i=\infty} [\mathrm{Inf}_\delta\, r(\xi_i, \delta) - r(\xi_i, \delta_0)] = 0 \tag{1.20}$$

where the symbol Inf_δ means infimum with respect to δ.

We shall say that a decision function δ is a Bayes solution in the strict sense if there exists an a priori distribution ξ such that δ is a Bayes solution relative to ξ. A decision function δ will be said to be a Bayes solution in the wide sense if there exists a sequence $\{\xi_i\}$ of a priori distributions such that δ is a Bayes solution relative to the sequence $\{\xi_i\}$.

One of the main results in Chapter 3 is that under very general conditions the class of all Bayes solutions in the wide sense is a complete class. It is also shown there that, under some further restrictions, the class of all Bayes solutions in the strict sense is already a complete class.

Consider the following simple example: Ω consists of two elements F_1 and F_2, where $F_i = \prod_{j=1}^{\infty} P_i(x_j)$ $(i = 1, 2)$, and $P_i(u)$ is a given one-dimensional distribution admitting a density function $p_i(u)$. The decision space D^t consists of two elements $d_1{}^t$ and $d_2{}^t$, where $d_i{}^t$ denotes the decision to accept the hypothesis that the true distribution F is equal to F_i $(i = 1, 2)$. Let g_i be the a priori probability that F_i is true $(i = 1, 2)$. Assume that experimentation is to be carried out in one stage and consists of m observations; i.e., the values of $X_1, \cdots, X_m$ are observed. Let the loss due to making a wrong terminal decision (accepting a hypothesis that is not true) be 1. Then a Bayes solution will be a decision function given as follows: After the sample $x_1, \cdots, x_m$ has been drawn, the a posteriori probability of the hypothesis H_i that F is equal to F_i is given by

$$g_{im} = \frac{g_i p_i(x_1) \cdots p_i(x_m)}{g_1 p_1(x_1) \cdots p_1(x_m) + g_2 p_2(x_1) \cdots p_2(x_m)} \qquad (i = 1, 2)$$

If $g_{1m} > g_{2m}$, accept H_1; if $g_{1m} < g_{2m}$, accept H_2; if $g_{1m} = g_{2m}$, any chance mechanisms may be used to decide between H_1 and H_2. The inequalities $g_{2m} \gtreqless g_{1m}$ are equivalent to $p_{2m}/p_{1m} \gtreqless g_1/g_2$, where $p_{im} = p_i(x_1) \cdots p_i(x_m)$. Thus the decision rule may be formulated as follows: If $p_{2m}/p_{1m} > g_1/g_2$, accept H_2; if $p_{2m}/p_{1m} < g_1/g_2$, accept H_1; if $p_{2m}/p_{1m} = g_1/g_2$, any chance mechanisms may be used to decide between H_1 and H_2. Let δ_c denote the above Bayes solution when $g_1/g_2 = c$. It follows from the results in Chapter 3 that the class of all Bayes solutions δ_c corresponding to all non-negative values of c is a complete class, provided that experimentation is restricted to a one-stage experiment with m observations.[7]

[7] For a more detailed discussion, see Section 5.1.1.

1.4.2 Decision Functions which Minimize the Maximum Risk (Minimax Solutions)

A decision function δ_0 is said to be a minimax solution of the decision problem if it minimizes the maximum of $r(F, \delta)$ with respect to F, i.e., if

$$\text{Sup}_F\, r(F, \delta_0) \leqq \text{Sup}_F\, r(F, \delta) \tag{1.21}$$

for all δ, where the symbol Sup_F stands for supremum with respect to F.

In the general theory of decision functions, as developed in Chapter 3, much attention is given to the theory of minimax solutions for two reasons: (1) a minimax solution seems, in general, to be a reasonable solution of the decision problem when an a priori distribution in Ω does not exist or is unknown to the experimenter; (2) the theory of minimax solutions plays an important role in deriving the basic results concerning complete classes of decision functions.

There is an intimate connection between minimax solutions and Bayes solutions. It will be seen in Chapter 3 that under general conditions a minimax solution is also a Bayes solution. More precisely, a minimax solution is, under some weak restrictions, a Bayes solution relative to a least favorable a priori distribution. An a priori distribution ξ_0 will be said to be least favorable if

$$\text{Inf}_\delta\, r(\xi_0, \delta) \geqq \text{Inf}_\delta\, r(\xi, \delta)$$

for all ξ.

In a number of cases a minimax solution can easily be obtained by finding an a priori probability measure ξ and a Bayes solution δ_ξ relative to ξ such that $\text{Sup}_F\, r(F, \delta_\xi) = r(\xi, \delta_\xi)$. Obviously δ_ξ is a minimax solution and ξ is a least favorable a priori distribution.

1.5 Relation to Earlier Theories

1.5.1 Testing a Hypothesis Viewed as a Special Case of the General Decision Problem

By a hypothesis we mean a statement that the unknown distribution F of X is an element of a given subset ω of Ω. For any non-empty subset ω of Ω, we shall use the symbol H_ω to denote the hypothesis that $F \epsilon \omega$. The problem of testing a hypothesis H is a special case of the general decision problem. In the case of testing a hypothesis H, the space D^t of terminal decisions consists of two elements ${d_1}^t$ and ${d_2}^t$, where ${d_1}^t$ denotes the decision to accept H and ${d_2}^t$ the decision to reject H.

The theories of testing hypotheses, as developed during the last

thirty years by Fisher, Neyman and Pearson, and their schools, deal almost exclusively with the case where experimentation is carried out in a single stage; i.e., it is determined in advance (before experimentation starts) how many and what kind of observations should be made during the whole course of experimentation. In other words, the choice of the experimenter is restricted to decision functions δ which satisfy the following condition: $\delta(D^e \mid 0) = 1$ and $\delta(D^t \mid y;\ s_1) = 1$ for any sequence y of real values and for any subset s_1 of the set of all positive integers.

It should also be remarked that the problem of design of experiments as treated by Fisher [18] and his school is contained as a special case in our formulation of the decision problem. If experimentation is carried out in a single stage, the problem of design reduces to deciding (in advance of the experimentation) how many and what kind of observations should be made during the whole course of experimentation. In other words, the problem of design reduces to the question of how to choose the value $\delta(0)$ of the decision function δ to be adopted. As an illustration, consider the following example: Suppose that we are interested in investigating the yields of m agricultural varieties $v_1, \cdots, v_m$. Suppose also that for the purpose of experimentation a piece of land consisting of m^2 plots $\{p_{ij}\}$ $(i, j = 1, \cdots, m)$ is available and that one variety can be planted in each plot p_{ij}. The problem of design that arises here is the problem of how to assign the varieties to the different plots. Let the chance variable X_{ijk} stand for the yield that would be produced on the plot p_{ij} if the variety v_k were assigned to it. Thus there are altogether m^3 possible chance variables X_{ijk} $(i, j, k = 1, \cdots, m)$. Since we can observe only m^2 of them (one variety is to be assigned to each plot), the problem of design here is simply the problem of which subset of m^2 chance variables of the set $\{X_{ijk}\}$ $(i, j, k = 1, \cdots, m)$ of m^3 chance variables we should select for observation. But this is precisely the problem of choosing the value $\delta(0)$ of the decision function δ to be adopted. A subset S of m^2 chance variables of the given set $\{X_{ijk}\}$ of m^3 chance variables is said to be a Latin square [8] if for any prescribed values of two of the three indices i, j, k there is precisely one element X_{ijk} in S having the prescribed values for these two indices. The solution of the design problem suggested by Fisher is to select a Latin square at random from the class of all possible Latin squares. Each Latin square is a particular element d^e of the space D^e. Let N be the total number of possible Latin squares. Then Fisher's solution of the design problem can be expressed in our notation and terminology as follows: We choose $\delta(0)$

[8] See, for example, Fisher [18].

to be the probability distribution in D^e for which $\delta(d^e \mid 0) = 1/N$ if d^e is a Latin square and $\delta(d^e \mid 0) = 0$ if d^e is not a Latin square.

It may be of interest to point out the relation between some of the notions in the present general decision theory (when applied to testing a hypothesis) and the corresponding notions in the Neyman-Pearson theory.[9] For this purpose we shall restrict ourselves to non-randomized decision functions according to which experimentation is carried out in a single stage by observing the values of the first N chance variables of the sequence $\{X_i\}$, since this is the only case treated in the Neyman-Pearson theory. In this case $\delta(0)$ assigns the probability 1 to the element $d^e = (1, 2, \cdots, N)$, and it is sufficient to define the value of $\delta(x; s)$ when s is equal to the set $(1, 2, \cdots, N)$. Thus, since δ is a non-randomized decision function and since D^t consists of the two elements $d_1{}^t$ and $d_2{}^t$, the decision function δ can be expressed by a function $d(x_1, \cdots, x_N)$, which is defined for all real numbers $x_1, \cdots, x_N$ and can take only the values $d_1{}^t$ and $d_2{}^t$ for each point $(x_1, \cdots, x_N)$. If $x_1, \cdots, x_N$ are the observed values of $X_1, \cdots, X_N$, respectively, then we accept the hypothesis H under test when $d(x_1, \cdots, x_N) = d_1{}^t$ and reject H when $d(x_1, \cdots, x_N) = d_2{}^t$. In the Neyman-Pearson theory the set of all sample points $x = (x_1, \cdots, x_N)$ for which we decide to reject H is called the critical region. Thus the choice of a critical region in the Neyman-Pearson theory is equivalent to the choice of a decision function in our terminology. Let the hypothesis H under test be the hypothesis that $F \epsilon \omega$. In the Neyman-Pearson theory, the probability that H will be rejected when some F, not an element of ω, is true is called the power of the critical region with respect to F. Thus the power function is a function of F defined for all F not in ω. The probability of rejecting H when some F is true that is an element of ω is called the size of the critical region with respect to F. Thus the size function is a function of F defined for all F in ω. The notions of size and power are special cases of the notion of risk in the general decision theory. In fact, let $W(F, d^t)$ be defined as follows: $W(F, d_1{}^t) = 0$ when $F \epsilon \omega$ and $= 1$ when $F \notin \omega$; $W(F, d_2{}^t) = 1$ when $F \epsilon \omega$ and $= 0$ when $F \notin \omega$. Thus $W(F, d^t)$ is a simple weight function. We can disregard the cost of experimentation here, since we restricted the choice of the experimenter to decision functions for which the expected cost of experimentation is the same constant. Then the simple risk corresponding to the above simple weight function is equal to the size of the critical region when $F \epsilon \omega$, and to (1 — power) when $F \notin \omega$.

In the Neyman-Pearson theory the choice of the critical region is subject to certain conditions imposed on the size function, such as

[9] See, for example, [35] and [37].

that the size function be equal to a prescribed constant α, or that the size function be bounded from above by a prescribed constant α. The imposition of such bounding conditions on some part of the risk function may be desirable when the errors due to possible wrong terminal decisions fall into classes which are of completely different kinds (e.g., one type of error may result in loss of life, the others in economic losses). The general decision theory, as developed in Chapter 3, remains applicable also when the choice of the decision function is subject to certain bounding conditions imposed on the probabilities of some types of errors. This is due to the fact that the class $\mathfrak{D}$ of decision functions δ to which the choice of the experimenter is restricted, is not assumed to be the class of *all* decision functions. The class $\mathfrak{D}$ is permitted to be any class satisfying a certain set of conditions. This set of conditions remains generally satisfied if bounding conditions of the above-mentioned type are imposed on the risk function.

In recent years a sequential method for testing a hypothesis H has been developed.[10] In this theory the restriction that the experiment is to be carried out in a single stage is removed. It is assumed, however, in that theory that (1) each stage of the experiment consists of a single observation, and (2) the chance variable X_i is observed in the ith stage. There is no loss of generality in the first restriction if we assume that the cost of experimentation depends on the total number of observations but not on the number of stages in which the experiment is carried out. The second restriction is more serious, since it does not leave freedom of choice for the selection of the chance variable to be observed at any stage of the experiment. In the special case when the chance variables $X_1, X_2, \cdots$, ad inf., are independently and identically distributed, there is no loss of generality in the second restriction either.

1.5.2 Point and Interval Estimation Viewed as Special Cases of the General Decision Problem

The problem of point estimation is the problem of deciding, on the basis of the results of the experiment, which element F of Ω should be adopted as our estimate of the true (but unknown) distribution of X. For any element F of Ω let $d_F{}^t$ denote the terminal decision to adopt F as our (point) estimate of the true distribution. Thus a point estimation problem is a special case of the general decision problem characterized by the fact that D^t consists of the elements $d_F{}^t$ corresponding to all F in Ω.

The theory of point estimation as developed by Fisher and others

[10] See, for example, [65].

during the last thirty years [11] deals almost exclusively with the case where experimentation is carried out in a single stage by observing the values of the first N chance variables in the sequence $\{X_i\}$. It may be of interest to point out the connection between some of the notions of the present general decision theory and the corresponding notions in the point estimation theory of Fisher and his school. For this purpose we shall restrict ourselves to non-randomized decision functions δ according to which experimentation is carried out in a single stage by observing the values of $X_1, \cdots, X_N$. In this case the non-randomized decision function δ can be expressed by a function $d(x_1, \cdots, x_N)$ defined for all real values $x_1, \cdots, x_N$. For any sample $(x_1, \cdots, x_N)$ the value of $d(x_1, \cdots, x_N)$ is an element d^t of D^t, and the experimenter makes the terminal decision $d(x_1, \cdots, x_N)$ when x_i is the observed value of X_i $(i = 1, \cdots, N)$. We shall also assume that Ω is a finite-parameter family of distribution functions F; i.e., each element F can be represented by particular values of a finite number of parameters, $\theta_1, \cdots, \theta_k$ (say). For the purpose of our discussion it will be sufficient to consider the case where there is only a single unknown parameter θ. We shall use the symbol $W(\theta, \theta^*)$ to denote $W(F, d_{F^*}{}^t)$, where θ is the parameter value corresponding to F and θ^* is the parameter value corresponding to F^*. Thus $W(\theta, \theta^*)$ is the loss suffered when θ is true and θ^* is adopted as a point estimate of θ. The decision function $d(x_1, \cdots, x_N)$ can be represented by a real-valued function $\theta^*(x_1, \cdots, x_N)$ such that the value $\theta^*(x_1, \cdots, x_N)$ is adopted as our point estimate of θ when $x_1, \cdots, x_N$ are the observed values of $X_1, \cdots, X_N$, respectively. In the theory of point estimation the function $\theta^*(x_1, \cdots, x_N)$ is also called an "estimator." Since the cost of experimentation is independent of the choice of the estimator $\theta^*(x_1, \cdots, x_N)$, we shall disregard it in computing the risk associated with the estimator $\theta^*(x_1, \cdots, x_N)$. Then the risk is simply the expected value of $W[\theta, \theta^*(x_1, \cdots, x_N)]$ when θ is true. Suppose that $W(\theta, \theta^*) = (\theta^* - \theta)^2$. Then the risk is simply the second moment of the estimator referred to the true parameter value θ. If $\theta^*(x_1, \cdots, x_N)$ is an unbiased estimator, i.e., if the expected value of $\theta^*(X_1, \cdots, X_N)$ is equal to the true parameter value θ, then the risk becomes equal to the variance of the estimator. A great deal of the literature on point estimation is devoted to the study of unbiased estimators with minimum variance, which are called efficient estimators. This theory can be regarded as a special case of the general decision theory when $W(\theta, \theta^*)$ is given by $(\theta^* - \theta)^2$. Minimum variance is not the only possible criterion for a "best" estimator. Various other definitions of

[11] See, for example, [15] and [16].

"best" estimators have been considered in the literature. Most of these theories can be represented as special cases of the general decision theory corresponding to some particular choices of the weight function $W(\theta, \theta^*)$. For example, Pitman [41] considered the problem of finding an estimator $\theta^*(x_1, \cdots, x_N)$ for which the probability that $|\theta^*(x_1, \cdots, x_N) - \theta| \leqq c$ is maximized, where c is a positive value.[12] This becomes equivalent to the problem of minimizing the risk if $W(\theta, \theta^*)$ is defined as follows: $W(\theta, \theta^*) = 0$ when $|\theta - \theta^*| \leqq c$, and $= 1$ when $|\theta - \theta^*| > c$.

The problem of interval estimation is again a special case of the general decision problem. For the purpose of the present discussion it will be sufficient to consider the case when Ω is a one-parameter family of distribution functions. Let θ be the unknown parameter. The problem of interval estimation may be formulated as follows: Let C be a given class of intervals. For example, C may be the class of all intervals, or the class of intervals of a given length, or the class of intervals whose length does not exceed a given value, and so on. The problem is to decide on the basis of the results of the experiment which element of C should be adopted as an interval estimate of θ. For any element I of C let $d_I{}^t$ denote the terminal decision to adopt I as an interval estimate of θ. Thus an interval estimation problem is a special case of the general decision problem where D^t consists of the elements $d_I{}^t$ corresponding to all elements I of a given class C of intervals.

In the theory of interval estimation as developed by Neyman [13] the only case considered is that where experimentation is carried out in a single stage by observing the first N chance variables of the sequence $\{X_i\}$. In this case any non-randomized decision function δ can be expressed by an interval function $I(x_1, \cdots, x_N)$ which associates an element I of C with each sample $(x_1, \cdots, x_N)$. The rule is then to take the terminal decision $d^t{}_{I(x_1, \cdots, x_N)}$ when $x_1, \cdots, x_N$ are the observed values. The weight function can now be represented as a function $W(\theta, I)$ of the true parameter value θ and the interval I adopted as an interval estimate of θ. We shall disregard the cost of experimentation, since it is independent of the choice of the decision rule when experimentation is carried out in one stage by observing the values of $X_1, \cdots, X_N$. Then the risk associated with a given interval estimator $I(x_1, \cdots, x_N)$ is simply the expected value of $W[\theta, I(X_1, \cdots, X_N)]$. A simple choice of $W(\theta, I)$ is to put $W(\theta, I) = 1$

[12] Pitman [41, page 401] calls an estimator θ^* "best" if the probability that $|\theta^* - \theta| \leqq c$ is maximized for all positive values c.

[13] See, for example, [38].

when $\theta \notin I$, and $= 0$ when $\theta \in I$. Then the risk associated with a given interval estimator $I(x_1, \cdots, x_N)$ is equal to the probability that $I(X_1, \cdots, X_N)$ will not cover the true parameter value. Neyman calls an interval function $I(x_1, \cdots, x_N)$ a confidence interval if the probability that $I(X_1, \cdots, X_N)$ will cover the true parameter value is equal to a fixed value γ, no matter what the true value of the parameter is. This fixed value γ is called the confidence coefficient associated with the confidence interval $I(x_1, \cdots, x_N)$. If the weight function $W(\theta, I)$ is defined as above, and if $I(x_1, \cdots, x_N)$ is a confidence interval with confidence coefficient γ, the risk associated with $I(x_1, \cdots, x_N)$ is equal to $1 - \gamma$.

1.6 Interpretation of the Decision Problem as a Zero Sum Two-Person Game

1.6.1 Definition of the Normalized Form of a Zero Sum Two-Person Game

The theory of games was developed by von Neumann [55]. It will be shown in Section 1.6.3 that the decision problem may be viewed as a zero sum two-person game and, therefore, the theory of such games can be applied to the statistical decision problem. A precise definition of a game was given by von Neumann. We shall not give it here, since for purposes of statistical applications it will be sufficient to consider merely the so-called normalized form of the game. As von Neumann has shown, any game can be brought into a normalized form.

The normalized form of a zero sum two-person game is given by von Neumann as follows: There are two players, and there is given a bounded and real-valued function $K(a, b)$ of two variables a and b, where a may be any point of a space A and b may be any point of a space B. Player 1 chooses a point a in A and player 2 chooses a point b in B, each choice being made in complete ignorance of the other. Player 1 then gets the amount $K(a, b)$ and player 2 the amount $-K(a, b)$. Clearly player 1 wishes to maximize $K(a, b)$ and player 2 wishes to minimize $K(a, b)$.

Any element a of A will be called a pure strategy of player 1, and any element b a pure strategy of player 2. A mixed strategy of player 1 is defined as follows: Instead of choosing a particular element a of A, player 1 chooses a probability measure ξ defined over a Borel field $\mathfrak{A}$ of subsets of A, and the point a is then selected by a chance mechanism constructed so that for any element α of $\mathfrak{A}$ the probability that the selected element a will be contained in α is equal to $\xi(\alpha)$. Similarly a mixed strategy of player 2 is given by a probability measure η defined

over a Borel field $\mathfrak{B}$ of subsets of B, and the element b is selected by a chance mechanism so that for any element β of $\mathfrak{B}$ the probability that the selected element b will be contained in β is equal to $\eta(\beta)$. The expected value of the outcome $K(a, b)$ is then given by

$$K^*(\xi, \eta) = \iint_{BA} K(a, b)\, d\xi\, d\eta \tag{1.22}$$

We can now reinterpret the value of $K(a, b)$ as the value of $K^*(\xi_a, \eta_b)$, where ξ_a and η_b are probability measures which assign probability 1 to a and b, respectively. In what follows, we shall write $K(\xi, \eta)$ for $K^*(\xi, \eta)$. Furthermore $K(a, b)$ will be used synonymously with $K(\xi_a, \eta_b)$, $K(a, \eta)$ synonymously with $K(\xi_a, \eta)$, and $K(\xi, b)$ synonymously with $K(\xi, \eta_b)$. This can be done without any danger of confusion. The function $K(\xi, \eta)$ is called the outcome function of the game.

1.6.2 Minimax, Minimal, Maximal, and Admissible Strategies

A strategy ξ_0 of player 1 will be said to be a minimax strategy if

$$\mathrm{Inf}_\eta\, K(\xi_0, \eta) \geqq \mathrm{Inf}_\eta\, K(\xi, \eta) \tag{1.23}$$

for all ξ. Similarly a strategy η_0 of player 2 will be said to be a minimax strategy if

$$\mathrm{Sup}_\xi\, K(\xi, \eta_0) \leqq \mathrm{Sup}_\xi\, K(\xi, \eta) \tag{1.24}$$

for all η. Minimax strategies play a fundamental role in the theory of two-person games, as we shall see in Chapter 2.

A strategy η_0 of player 2 is said to be minimal relative to a strategy ξ of player 1 if

$$K(\xi, \eta_0) = \mathrm{Min}_\eta\, K(\xi, \eta) \tag{1.25}$$

Clearly, if η_0 is a minimal strategy relative to ξ, then η_0 is an optimum strategy for player 2 when player 1 uses the strategy ξ.

A strategy ξ_0 of player 1 will be said to be maximal relative to a strategy η of player 2 if

$$K(\xi_0, \eta) = \mathrm{Max}_\xi\, K(\xi, \eta) \tag{1.26}$$

If player 2 uses the strategy η, then ξ_0 is an optimum strategy for player 1.

A strategy η_0 of player 2 will be said to be minimal relative to the sequence $\{\xi_i\}$ $(i = 1, 2, \cdots)$ of strategies of player 1 if

$$\lim_{i=\infty} [K(\xi_i, \eta_0) - \operatorname{Inf}_\eta K(\xi_i, \eta)] = 0 \tag{1.27}$$

A maximal strategy ξ_0 of player 1 relative to a sequence $\{\eta_i\}$ of strategies of player 2 is defined similarly.

A strategy η will be said to be minimal in the strict sense if there exists a strategy ξ of player 1 such that η is minimal relative to ξ. A strategy η will be said to be minimal in the wide sense if there exists a sequence $\{\xi_i\}$ such that η is minimal relative to $\{\xi_i\}$. Maximal strategies ξ in the strict and wide sense are defined correspondingly.

A strategy η_1 of player 2 is said to be uniformly better than the strategy η_2 if

$$K(\xi, \eta_1) \leqq K(\xi, \eta_2) \tag{1.28}$$

for all ξ, and if

$$K(\xi, \eta_1) < K(\xi, \eta_2) \tag{1.29}$$

for at least one ξ. Similarly a strategy ξ_1 of player 1 is said to be uniformly better than the strategy ξ_2 if

$$K(\xi_1, \eta) \geqq K(\xi_2, \eta) \tag{1.30}$$

for all η, and if

$$K(\xi_1, \eta) > K(\xi_2, \eta) \tag{1.31}$$

for at least one η.

A strategy of player i $(i = 1, 2)$ is said to be admissible if there is no uniformly better strategy for player i.

A class C of strategies of player i $(i = 1, 2)$ will be said to be complete if for any strategy not in C there exists a strategy in C that is uniformly better.

1.6.3 The Decision Problem Viewed as a Zero Sum Two-Person Game

In a decision problem the experimenter wishes to minimize the risk $r(F, \delta)$. The risk, however, depends on two variables F and δ, and the experimenter can choose only the decision function δ but not the true distribution F. The true distribution F, we may say, is chosen by Nature, and Nature's choice is unknown to the experimenter. Thus the situation that arises here is very similar to that of a two-person game. As a matter of fact, the decision problem can be interpreted as a zero sum two-person game by setting up the following correspondence.

Two-Person Game	Decision Problem
Player 1	Nature
Player 2	Experimenter
Pure strategy a of player 1	Choice of true distribution F by Nature
Space A of pure strategies of player 1	Space Ω
Pure strategy b of player 2	Choice of decision function δ by experimenter
Space B of pure strategies of player 2	Space $\mathfrak{D}$ of all possible decision functions δ
Outcome $K(a, b)$	Risk $r(F, \delta)$
Mixed strategy ξ of player 1	A priori distribution ξ in Ω
Mixed strategy η of player 2	Probability measure η defined over a Borel field of subsets of the space $\mathfrak{D}$
Outcome $K(\xi, \eta)$	$r(\xi, \eta) = \int_{\mathfrak{D}} \int_{\Omega} r(F, \delta)\, d\xi\, d\eta$
Minimax strategy of player 2	Minimax solution of decision problem
Minimax strategy of player 1	Least favorable a priori distribution in Ω
Minimal strategy of player 2	Bayes solution
Admissible strategy of player 2	Admissible decision function

It would have been possible to regard only the non-randomized decision functions as the pure strategies of the experimenter. The choice of a probability measure (mixed strategy) in the space of all non-randomized decision functions can be shown to be equivalent to the choice of some randomized decision function δ. For purposes of developing the general theory, as given in Chapter 3, it seemed, however, to be more convenient to regard the randomized decision functions themselves as the pure strategies. By doing so, it will be possible to disregard altogether mixed strategies for the experimenter, since, as will be seen in Chapter 3, the choice of a probability measure η in the space $\mathfrak{D}$ is equivalent to the choice of a particular element δ of $\mathfrak{D}$.

The analogy between the decision problem and a two-person game seems to be complete, except for one point. Whereas the experimenter wishes to minimize the risk $r(F, \delta)$, we can hardly say that Nature wishes to maximize $r(F, \delta)$. Nevertheless, since Nature's choice is unknown to the experimenter, it is perhaps not unreasonable for the experimenter to behave as if Nature wanted to maximize the risk. But, even if one is not willing to take this attitude, the theory of games remains of fundamental importance for the problem of statistical decisions, since, as will be seen in Chapter 3, it leads to basic results concerning admissible decision functions and complete classes of decision functions.

The theory of zero sum two-person games was developed by von Neumann for finite spaces A and B, i.e., when both players have only a finite number of pure strategies at their disposal. In statistical decision problems, however, the corresponding spaces Ω and $\mathfrak{D}$ generally have infinitely many elements. In the next chapter the theory of zero sum two-person games is extended to the case where the players have infinitely many strategies at their disposal.

1.7 Note on Some Ideas and Results Preceding the Present Developments

Until about ten years ago, the available statistical theories, except for a few scattered results, were restricted in two important directions: (1) only decision functions were treated for which experimentation is carried out in a single stage; (2) the decision problems were restricted to problems of testing a hypothesis, and that of point and interval estimation.

Among the few early results not subject to restriction (1), a double sampling inspection procedure by Dodge and Romig [14] may be mentioned. According to their scheme the decision whether or not a second sample should be drawn before a terminal decision is made depends on the outcome of the observations in the first sample. The need for multi-stage experimentation had been recognized long before any systematic theory regarding such experimentation was available. This was clearly shown by the occasional practice in the past of designing a large scale experiment in successive stages. A very interesting example of this type is the series of sample censuses of area of jute in Bengal carried out under the direction of Mahalanobis [31]. A number of preliminary sample censuses were taken, and the information contained in these samples was then used to design the final sampling of the whole jute area.

The possibility of an extension of the theory of testing a hypothesis H by admitting three terminal decisions, acceptance of H, rejection of H, and no choice between H and non-H, was considered by Neyman and Pearson [34] as early as 1933. The "decision" character of the test and estimation procedures has been emphasized by Neyman, who termed the adoption of a particular test or estimation procedure "inductive behavior." [14]

The basic ideas of a general theory of non-randomized decision functions when experimentation is carried out in a single stage and when the space D^t of terminal decisions is any general space were first out-

[14] See, for example, [38].

lined by the author in a publication in 1939 [56]. In this publication the notions of weight and risk functions are introduced [15] and the nature of the minimax and Bayes solutions are studied. The results of this paper were considerably extended, and the relationship to the theory of games was recognized in 1945 [59], but the assumption of one-stage experimentation had still been maintained.

A major advance in the theory of multi-stage experimentation took place during World War II with the development of sequential analysis.[16] This theory deals mainly with the problem of testing a hypothesis (D^t contains only two elements) with no definite upper bound on the number of stages of experimentation. It is assumed, however, that the ith stage of experimentation consists of a single observation on X_i ($i = 1, 2, \cdots$, ad inf.). Thus, if the experiment is carried out in n stages, it consists of the observations on $X_1, \cdots, X_n$. The number of stages of the experiment is, of course, a chance variable, since it depends on the observed values obtained. The main part of the theory consists of the development of the so-called sequential probability ratio test, a particular sequential method for testing a hypothesis. Contributions to the further development of sequential analysis have been made in the last few years by several authors in this country and in England, notably by Anscombe [2], Armitage [3], Barnard [6], Bartlett [8], Blackwell [10–12], Burman [13], Girshick [19–21], Mosteller [21], Savage [46], Stein [49–52], Stockman [53], Wald [57, 60–71], and Wolfowitz [69, 71, 73–75].

A very interesting paper by Bartky in 1943 [7] may be regarded as a forerunner of sequential analysis. In this paper a multiple sampling scheme is given for testing the mean of a binomial distribution.

In 1945 Stein [49] published a highly interesting double sampling method for obtaining a confidence interval of fixed length for the mean of a normal distribution with unknown variance. His method is particularly interesting, since no confidence interval of fixed length can be obtained with any single sampling method.

The concept of a complete class of decision functions was introduced by Lehmann, and the first result regarding such classes is due to him [30]. He obtained the minimal complete class of decision functions in the following special case: the chance variables $X_1, X_2, \cdots, X_n$ admit a

[15] The idea of assigning weights to the various possible wrong decisions had already been considered by Neyman and Pearson as early as 1933 [34]. Also the minimax principle is mentioned in [34] as a possible approach to the decision problem.

[16] See, for example, [48b] and [65].

joint probability density function $f(x_1, \cdots, x_n, \theta)$ which is known except for the value of a single parameter θ (Ω is a one-parameter family of distribution functions). Experimentation is carried out in a single stage by observing the values of $X_1, \cdots, X_n$. The function $f(x_1, \cdots, x_n, \theta)$ satisfies essentially the conditions formulated by Neyman [37] to insure the existence of a uniformly most powerful unbiased test [these include the fulfillment of a certain differential equation by the function $f(x_1, \cdots, x_n, \theta)$]. The problem considered is to test the hypothesis that θ is equal to a specified value θ_0.

Soon after Lehmann's paper appeared, the author obtained general results concerning complete classes of decision functions in three successive papers [66, 67, 70], the first of which deals with the non-sequential case and the second and the third with the sequential case. It was shown that under very general conditions the class of all Bayes solutions is a complete class.

The general theory of non-sequential decision functions contained in the author's paper in 1945 [59] was extended to the sequential case in two successive papers in 1947 [67] and 1949 [70]. These papers deal with the general decision problem where experimentation may be carried out in any number of stages, but it is assumed that the ith stage of the experiment consists of a single observation on X_i.

Stein [52] was the first to formulate a model for statistical decision procedures which includes the design of experimentation (selection of the chance variables to be observed) as a part of the decision problem. His scheme is, however, restricted in several ways. The space Ω and the space D^t of terminal decisions are assumed to be finite.[17] Furthermore there is a fixed finite upper bound for the total number of observations that can be made. The problem considered by Stein is related to, but different from, and more special than, the problem treated in the present book. He is concerned with the problem of finding a decision function which is optimum in the sense that, under some side conditions on the probabilities for making wrong decisions, it minimizes the expected cost of experimentation when a particular element F_0 of Ω is the true distribution. His main result consists in giving sufficient conditions for a decision function to be optimum in his sense. The question whether decision functions satisfying his sufficient conditions always exist is left open. In a number of special cases, however, he verified that such decision functions exist.

The present book is mainly an outgrowth of several previous publica-

[17] Actually it is not assumed that Ω is finite, but the theory developed by Stein is such that only a finite number of elements of Ω enter and the rest of the space Ω can be disregarded.

tions of the author on the general theory of decision functions [59, 66, 67, 70], and it contains a considerable expansion and generalization of the ideas and results obtained in these papers. Particularly, the restriction is dropped that the ith stage of the experiment consists of a single observation on X_i, making it possible to treat the design of experimentation as a part of the decision problem.

Chapter 2. ZERO SUM TWO-PERSON GAMES WITH INFINITELY MANY STRATEGIES

2.1 Conditions for Strict Determinateness of a Game

2.1.1 The Problem of Strict Determinateness of a Game and the Introduction of an Intrinsic Metric

Extending von Neumann's definition for finite spaces of strategies to the infinite case,[1] we shall say that a game is strictly determined if

$$\text{Sup}_\xi \text{Inf}_\eta K(\xi, \eta) = \text{Inf}_\eta \text{Sup}_\xi K(\xi, \eta) \tag{2.1}$$

where the symbol Sup_ξ stands for supremum with respect to ξ and Inf_η stands for infimum with respect to η. The common value of the left- and right-hand members of (2.1) is called the value of the game. The question of strict determinateness is of basic importance in the theory of games for the following reason. If the game is strictly determined and both players use minimax strategies, provided that such strategies exist, then neither player can improve his situation by finding out his opponent's strategy; i.e., neither player will have any inducement to abandon his own minimax strategy even if he finds out his opponent's strategy. Thus, for strictly determined games, the use of minimax strategies creates a perfectly stable situation and the minimax strategies may be regarded as good strategies. On the other hand, if (2.1) does not hold, no stable situation exists; i.e., no matter what strategies are chosen by the players, at least one of them can improve his situation by finding out his opponent's strategy.

The main theorem proved by von Neumann [2] states that, if the spaces A and B of pure strategies are finite, (2.1) always holds; i.e., the game is always strictly determined. A game with infinitely many strategies, however, is not necessarily strictly determined, as shown by the following simple example. Let A and B each be the space of all positive integers. The outcome $K(a, b) = 1$ if $a > b$, $= 0$ if $a = b$, and $= -1$ if $a < b$. One can easily verify that for this game we have $\text{Sup}_\xi \text{Inf}_\eta K(\xi, \eta) = -1$ and $\text{Inf}_\eta \text{Sup}_\xi K(\xi, \eta) = 1$. Thus (2.1) does not hold.

Necessary and sufficient conditions for strict determinateness were given by the author [58] for the case when the spaces A and B have

[1] See Section 14.5.1 in [55].

[2] See Section 17.6 of [55].

countably many elements. A special result for spaces A and B with continuously many elements was obtained by Ville [54]. He considered the case where A and B are finite and closed intervals of the real axis and $K(a, b)$ is a continuous function of a and b. He proved that in this case the game is strictly determined. A more general result was obtained later by the author [67]. Conditions for strict determinateness will be given here in Sections 2.1.3, 2.1.4, and 2.1.5. The results contained in Sections 2.1.4 and 2.1.5 go somewhat beyond previously published results.

To give a precise meaning to the relation (2.1) in the case of infinite spaces A and B, we have to define the Borel field $\mathfrak{A}$ of subsets of A and the Borel field $\mathfrak{B}$ of subsets of B for which the probability measures ξ and η are defined, respectively. We shall define Borel fields $\mathfrak{A}$ and $\mathfrak{B}$ with the help of an intrinsic metric in the spaces A and B. The (intrinsic) distance $\delta(a_1, a_2)$ of two elements a_1 and a_2 is defined by [3]

$$\delta(a_1, a_2) = \mathrm{Sup}_b \left| K(a_1, b) - K(a_2, b) \right| \tag{2.2}$$

Similarly the (intrinsic) distance of two elements b_1 and b_2 in B is defined by

$$\delta(b_1, b_2) = \mathrm{Sup}_a \left| K(a, b_1) - K(a, b_2) \right| \tag{2.3}$$

The metric in A, as well as that in B, satisfies the triangle inequality,[4] but it may happen that two different elements of A, or B, have the distance zero. We can, however, replace the original spaces A and B by the spaces A^* and B^* defined as follows: For any element a of A let α_a be the set of all elements of A which have the distance zero from a. Clearly, for any two elements a_1 and a_2 of A, the sets α_{a_1} and α_{a_2} are either disjoint or identical. The space A^* is then the space of all subsets α_a of A. Let a_1^* and a_2^* be two different elements of A^*. Then there exist two elements a_1 and a_2 of A such that $a_1^* = \alpha_{a_1}$, $a_2^* = \alpha_{a_2}$, and α_{a_1} has no common element with α_{a_2}. We put $\delta(a_1^*, a_2^*) = \delta(a_1, a_2)$. The space B^* and the metric in B^* are defined in a similar way. The distance between two different elements of A^* or B^* is always positive. In the theory of games only the spaces A^* and B^* play a relevant role. In what follows we shall assume that any two different elements of A or B have a positive distance. There is no loss of generality in this assumption, since the spaces A and B can

[3] A similar distance definition corresponding to a certain function of two variables was used by Helly [24] in connection with linear spaces. He refers to it as the "polar distance function."

[4] The triangle inequality is said to be satisfied if for any three points a_1, a_2, and a_3 of the space we have $\delta(a_1, a_2) + \delta(a_2, a_3) \geqq \delta(a_1, a_3)$.

always be replaced by A^* and B^*, respectively. Thus the distance definitions given in (2.2) and (2.3) make the spaces A and B metric spaces.

The distance definitions given in (2.2) and (2.3) can be extended to the spaces of mixed strategies. We put

$$\delta(\xi_1, \xi_2) = \mathrm{Sup}_\eta \left| K(\xi_1, \eta) - K(\xi_2, \eta) \right| \tag{2.4}$$

and

$$\delta(\eta_1, \eta_2) = \mathrm{Sup}_\xi \left| K(\xi, \eta_1) - K(\xi, \eta_2) \right| \tag{2.5}$$

Of particular interest are the Borel fields $\mathfrak{A}_1$ and $\mathfrak{B}_1$, where $\mathfrak{A}_1$ is the smallest Borel field of subsets of A containing all open subsets, in the sense of the metric (2.2), of A as elements, and $\mathfrak{B}_1$ is the smallest Borel field of subsets of B containing all open subsets of B as elements. Clearly all denumerable subsets of A and B are elements of $\mathfrak{A}_1$ and $\mathfrak{B}_1$, respectively. It will be seen in Section 2.1.4 that, if the space $A(B)$ is separable [4a] in the sense of its intrinsic metric, there will be little interest in considering a Borel field $\mathfrak{A}(\mathfrak{B})$ different from $\mathfrak{A}_1(\mathfrak{B}_1)$. However, for non-separable spaces of strategies, the consideration of Borel fields different from $\mathfrak{A}_1$ and $\mathfrak{B}_1$ may be useful, as Section 2.1.5 will indicate. Whenever we speak of a subset of $A(B)$, we shall always mean an element of the Borel field $\mathfrak{A}(\mathfrak{B})$.

Let $\mathfrak{A}_0$ be the smallest Borel field containing all denumerable subsets of A as elements, and let $\mathfrak{B}_0$ be the smallest Borel field containing all denumerable subsets of B as elements. We shall consider only Borel fields $\mathfrak{A}$ and $\mathfrak{B}$ which contain $\mathfrak{A}_0$ and $\mathfrak{B}_0$, respectively, as subfields.

Any theorem or lemma stated in the present chapter is meant to be valid for $\mathfrak{A} = \mathfrak{A}_1$ and $\mathfrak{B} = \mathfrak{B}_1$ unless stated otherwise.

Let $C = A \times B$ be the Cartesian product of A and B,[5] and let $\mathfrak{C}$ be the smallest Borel field of subsets of C which contains the Cartesian product of any member of $\mathfrak{A}$ with any member of $\mathfrak{B}$. In this study we shall restrict ourselves to games for which the outcome $K(a, b)$ is a bounded function of a and b and is measurable ($\mathfrak{C}$). In the next section we shall prove some lemmas which will then be used to derive conditions for strict determinateness of a game.

It is of interest to note that if $\mathfrak{A} = \mathfrak{A}_1$, $\mathfrak{B} = \mathfrak{B}_1$, and one of the spaces A and B is separable, $K(a, b)$ is always measurable ($\mathfrak{C}$). For example, let A be separable and let γ be the subset of C consisting of all points (a, b) for which $K(a, b) < r$, where r is a given real number. We shall now show that γ is a member of $\mathfrak{C}$. Let α be the subset of A consisting of every point a for which there exists an element b of B

[4a] For a definition of "separable," see Section 2.1.4.

[5] See page 82 of [44].

such that (a, b) is an element of γ. Clearly α is an open subset of A. For any positive value ρ, let $S(a, \rho)$ denote the closed sphere in A with center a and radius ρ; i.e., $S(a, \rho)$ is the totality of all points a' whose distance from a does not exceed ρ. For any subset α' of α let $\beta(\alpha')$ be the totality of all those points b of B for which the Cartesian product $\alpha' \times b$ is a subset of γ. Clearly for any a and ρ for which $S(a, \rho) \subset \alpha$, the set $\beta[S(a, \rho)]$ is open. Let α^* be a denumerable dense subset of α, and consider the subset γ^* of C given by

$$\gamma^* = \sum_{a,\rho} \{S(a, \rho) \times \beta[S(a, \rho)]\}$$

where the summation is to be taken over all pairs (a, ρ) for which $a \epsilon \alpha^*$, ρ is rational, and $S(a, \rho)$ is a subset of α. Since for each pair (a, ρ), the set $S(a, \rho) \times \beta[S(a, \rho)]$ is a member of $\mathfrak{C}$, the set γ^* is also a member of $\mathfrak{C}$. Clearly γ^* is a subset of γ. We shall now show that $\gamma^* = \gamma$. Let (a_0, b_0) be any point of γ. We merely have to show that (a_0, b_0) is a point of γ^*. Clearly there exists a positive value ρ_0 such that $S(a_0, \rho_0) \times S(b_0, \rho_0)$ is a subset of γ, where $S(b, \rho)$ denotes the closed sphere in B with center b and radius ρ. Hence there exists an element a_1 of α^* and a positive rational number ρ_1 such that $S(a_1, \rho_1) \subset S(a_0, \rho_0)$ and a_0 is an element of $S(a_1, \rho_1)$. Clearly $S(a_1, \rho_1) \times \beta[S(a_1, \rho_1)]$ contains the point (a_0, b_0). Since $S(a_1, \rho_1) \times \beta[S(a_1, \rho_1)]$ is a subset of γ^*, the point (a_0, b_0) must be an element of γ^*. This completes the proof of our statement that $K(a, b)$ is measurable $(\mathfrak{C})$ where one of the spaces A and B is separable.

2.1.2 Some Lemmas

In what follows, for any subset α of A the symbol ξ_α will denote a probability measure ξ on A for which $\xi(\alpha) = 1$. Similarly, for any subset β of B, η_β will denote a probability measure η on B for which $\eta(\beta) = 1$. We shall now prove the following lemma.

Lemma 2.1. Let $\{\alpha_i\}$ $(i = 1, 2, \cdots,$ *ad inf.*$)$ *be a sequence of subsets of* A *such that* $\alpha_i \subset \alpha_{i+1}$ *and let* $\alpha = \sum_{i=1}^{\infty} \alpha_i$. *Then*

$$\lim_{i=\infty} \mathrm{Sup}_{\xi_{\alpha_i}} \mathrm{Inf}_\eta K(\xi_{\alpha_i}, \eta) = \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_\eta K(\xi_\alpha, \eta) \tag{2.6}$$

Proof: Clearly the limit as $i \to \infty$ of $\mathrm{Sup}_{\xi_{\alpha_i}} \mathrm{Inf}_\eta K(\xi_{\alpha_i}, \eta)$ exists and cannot exceed the value of the right-hand member of (2.6). Put

$$\lim_{i=\infty} \mathrm{Sup}_{\xi_{\alpha_i}} \mathrm{Inf}_\eta K(\xi_{\alpha_i}, \eta) = \rho \tag{2.7}$$

and

$$\mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_\eta K(\xi_\alpha, \eta) = \rho + \delta \quad (\delta \geqq 0) \tag{2.8}$$

Suppose that $\delta > 0$. Then there exists a probability measure $\xi_\alpha{}^0$ such that

$$K(\xi_\alpha{}^0, \eta) \geqq \rho + \frac{\delta}{2} \tag{2.9}$$

for all η. Let $\xi_{\alpha_i}{}^0$ be the probability measure given as follows: For any subset α^* of α_i we have

$$\xi_{\alpha_i}{}^0(\alpha^*) = \frac{\xi_\alpha{}^0(\alpha^*)}{\xi_\alpha{}^0(\alpha_i)}. \tag{2.10}$$

Then, since $\lim_{i=\infty} \xi_\alpha{}^0(\alpha - \alpha_i) = 0$ and since $K(a, b)$ is uniformly bounded, we have

$$\lim_{i=\infty} K(\xi_{\alpha_i}{}^0, \eta) = K(\xi_\alpha{}^0, \eta) \tag{2.11}$$

uniformly in η. Hence for sufficiently large i the inequality

$$\mathrm{Inf}_\eta\, K(\xi_{\alpha_i}{}^0, \eta) \geqq \rho + \frac{\delta}{3} \tag{2.12}$$

holds. But this is not possible because of (2.7). Thus, $\delta = 0$ and Lemma 2.1 is proved.

Interchanging the role of the two players, Lemma 2.1 yields the following lemma.

Lemma 2.2. Let $\{\beta_i\}$ $(i = 1, 2, \cdots,$ *ad inf.*$)$ *be a sequence of subsets of* B *such that* $\beta_i \subset \beta_{i+1}$ *and let* $\beta = \sum_{i=1}^{\infty} \beta_i$. *Then*

$$\lim_{i=\infty} \mathrm{Inf}_{\eta_{\beta_i}} \mathrm{Sup}_\xi\, K(\xi, \eta_{\beta_i}) = \mathrm{Inf}_{\eta_\beta} \mathrm{Sup}_\xi\, K(\xi, \eta_\beta) \tag{2.13}$$

We shall now prove the following lemma.

Lemma 2.3. The inequality

$$\mathrm{Sup}_\xi \mathrm{Inf}_\eta\, K(\xi, \eta) \leqq \mathrm{Inf}_\eta \mathrm{Sup}_\xi\, K(\xi, \eta) \tag{2.14}$$

always holds.

Proof: Clearly

$$K(\xi, \eta) \leqq \mathrm{Sup}_\xi\, K(\xi, \eta) \tag{2.15}$$

Hence

$$\mathrm{Inf}_\eta\, K(\xi, \eta) \leqq \mathrm{Inf}_\eta \mathrm{Sup}_\xi\, K(\xi, \eta) \tag{2.16}$$

This gives

$$\mathrm{Sup}_\xi \mathrm{Inf}_\eta\, K(\xi, \eta) \leqq \mathrm{Inf}_\eta \mathrm{Sup}_\xi\, K(\xi, \eta) \tag{2.17}$$

and our lemma is proved. This proof is essentially the same as that given by von Neumann [55] for finite spaces A and B.

Lemma 2.4. If there exists a subset α of A such that

$$\text{Sup}_{\xi_\alpha} \text{Inf}_\eta K(\xi_\alpha, \eta) = \text{Inf}_\eta \text{Sup}_\xi K(\xi, \eta) \tag{2.18}$$

then the game is strictly determined.

Proof: Suppose that there exists a subset α of A for which (2.18) holds. Clearly

$$\text{Sup}_\xi \text{Inf}_\eta K(\xi, \eta) \geqq \text{Sup}_{\xi_\alpha} \text{Inf}_\eta K(\xi_\alpha, \eta) \tag{2.19}$$

From this and (2.18) we obtain

$$\text{Sup}_\xi \text{Inf}_\eta K(\xi, \eta) \geqq \text{Inf}_\eta \text{Sup}_\xi K(\xi, \eta) \tag{2.20}$$

From Lemma 2.3 it follows that the equality sign must hold in (2.20), and Lemma 2.4 is proved.

Interchanging the two players, Lemma 2.4 yields the following lemma.

Lemma 2.5. If there exists a subset β of B such that

$$\text{Inf}_{\eta_\beta} \text{Sup}_\xi K(\xi, \eta_\beta) = \text{Sup}_\xi \text{Inf}_\eta K(\xi, \eta) \tag{2.21}$$

then the game is strictly determined.

2.1.3 The Case when the Space of Strategies of One of the Players Is Conditionally Compact

We shall consider here the case when one of the spaces A or B is conditionally compact in the sense of its intrinsic metric given in (2.2) or (2.3). A metric space C is said to be conditionally compact if any sequence $\{c_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of elements of C admits a Cauchy subsequence $\{c_{i_j}\}$ $(j = 1, 2, \cdots,$ ad inf.), i.e., a subsequence $\{c_{i_j}\}$ with the property that

$$\lim_{j_1, j_2 = \infty} \delta(c_{i_{j_1}}, c_{i_{j_2}}) = 0$$

where $\delta(c_k, c_l)$ denotes the distance of the points c_k and c_l.

Theorem 2.1. If one of the spaces A and B is conditionally compact, both spaces are conditionally compact.

Proof: Suppose that A is conditionally compact. Then for any $\epsilon_1 > 0$ there exists a finite subset α of A that is ϵ_1-dense in A. A subset α of A is said to be ϵ-dense in A if for any point a in A there exists a point a' in α such that $\delta(a, a') \leqq \epsilon$. If we replace the space A by its ϵ_1-dense subset α, then the metric in the space B, as defined in (2.3),

will be changed. Let $\delta_\alpha(b_1, b_2)$ denote the new distance when A is replaced by α. Since α is finite, it follows easily from the definition of $\delta_\alpha(b_1, b_2)$ that for any $\epsilon_2 > 0$ there exists a finite subset $\beta(\epsilon_2)$ of B which is ϵ_2-dense in B in the sense of the metric $\delta_\alpha(b_1, b_2)$. Clearly

$$|\delta(b_1, b_2) - \delta_\alpha(b_1, b_2)| \leqq 2\epsilon_1 \tag{2.22}$$

Hence the set $\beta(\epsilon_2)$ must be $(\epsilon_2 + 2\epsilon_1)$-dense in B in the sense of the original metric $\delta(b_1, b_2)$. Since ϵ_1 and ϵ_2 can be chosen arbitrarily small, we find that for any $\delta > 0$ there exists a finite subset β of B that is δ-dense in B according to the original metric $\delta(b_1, b_2)$. But this is equivalent to conditional compactness,[6] and Theorem 2.1 is proved.

Theorem 2.2. If one of the spaces A and B is conditionally compact, the game is strictly determined.

Proof: Suppose that A is conditionally compact. Then, according to Theorem 2.1, B is also conditionally compact. Let ϵ be any positive value. Because of the conditional compactness of the spaces A and B we can subdivide A and B into a finite number of non-empty disjoint subsets the diameter of each of which does not exceed ϵ. Let $A_1, \cdots, A_k$ be non-empty subsets of A, and $B_1, \cdots, B_l$ non-empty subsets of B satisfying the above conditions; i.e.,

$$A_1 + \cdots + A_k = A; \qquad B_1 + \cdots + B_l = B \tag{2.23}$$

the sets $A_1, \cdots, A_k, B_1, \cdots, B_l$ are disjoint; and the diameter of any of these sets does not exceed ϵ. Let a_i be a particular point of A_i $(i = 1, \cdots, k)$ and b_j a particular point of B_j $(j = 1, \cdots, l)$, and let α denote the finite subset $\{a_1, \cdots, a_k\}$ of A and β the finite subset $\{b_1, \cdots, b_l\}$ of B. With any probability measure ξ^0 on A we associate the probability measure $\xi_\alpha{}^0$ defined as follows: $\xi_\alpha{}^0(a_i) = \xi^0(A_i)$ $(i = 1, 2, \cdots, k)$. Similarly, with any probability measure η^0 on B, we associate the probability measure $\eta_\beta{}^0$ given by $\eta_\beta{}^0(b_j) = \eta^0(B_j)$ $(j = 1, 2, \cdots, l)$. We then have

$$|K(\xi^0, \eta) - K(\xi_\alpha{}^0, \eta)| \leqq \epsilon \tag{2.24}$$

for all η, and

$$|K(\xi, \eta^0) - K(\xi, \eta_\beta{}^0)| \leqq \epsilon \tag{2.25}$$

for all ξ. It follows easily from (2.24) that

$$\begin{aligned} \mathrm{Sup}_\xi \mathrm{Inf}_\eta K(\xi, \eta) - \epsilon &\leqq \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_\eta K(\xi_\alpha, \eta) \\ &\leqq \mathrm{Sup}_\xi \mathrm{Inf}_\eta K(\xi, \eta) \end{aligned} \tag{2.26}$$

[6] See, for example, page 108 of [23].

From (2.25) we obtain

$$\text{(2.27)} \qquad \begin{aligned} \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_{\eta} K(\xi_\alpha, \eta) &\leqq \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_{\eta_\beta} K(\xi_\alpha, \eta_\beta) \\ &\leqq \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_{\eta} K(\xi_\alpha, \eta) + \epsilon \end{aligned}$$

Equations (2.26) and (2.27) imply that

$$\text{(2.28)} \qquad \begin{aligned} \mathrm{Sup}_{\xi} \mathrm{Inf}_{\eta} K(\xi, \eta) - \epsilon &\leqq \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_{\eta_\beta} K(\xi_\alpha, \eta_\beta) \\ &\leqq \mathrm{Sup}_{\xi} \mathrm{Inf}_{\eta} K(\xi, \eta) + \epsilon \end{aligned}$$

In a similar way, we obtain the inequality

$$\text{(2.29)} \qquad \begin{aligned} \mathrm{Inf}_{\eta} \mathrm{Sup}_{\xi} K(\xi, \eta) - \epsilon &\leqq \mathrm{Inf}_{\eta_\beta} \mathrm{Sup}_{\xi_\alpha} K(\xi_\alpha, \eta_\beta) \\ &\leqq \mathrm{Inf}_{\eta} \mathrm{Sup}_{\xi} K(\xi, \eta) + \epsilon \end{aligned}$$

According to von Neumann's theorem, for finite spaces we have

$$\text{(2.30)} \qquad \mathrm{Sup}_{\xi_\alpha} \mathrm{Inf}_{\eta_\beta} K(\xi_\alpha, \eta_\beta) = \mathrm{Inf}_{\eta_\beta} \mathrm{Sup}_{\xi_\alpha} K(\xi_\alpha, \eta_\beta)$$

Since ϵ can be chosen arbitrarily small, Theorem 2.2 follows from (2.28), (2.29), and (2.30).

In the remainder of this section we shall prove some theorems concerning the change in the value of the game when the original spaces A and B are replaced by some subsets α and β, respectively. In what follows, for any subset α of A and any subset β of B, we shall mean by the game relative to (α, β) the game we obtain when A is replaced by α and B by β. The Borel fields $\mathfrak{A}$ and $\mathfrak{B}$ will be assumed to remain unchanged when A and B are replaced by α and β, respectively. Thus the replacement of $A(B)$ by $\alpha(\beta)$ simply means that player 1(2) can use any probability measures $\xi(\eta)$ defined over the elements of $\mathfrak{A}(\mathfrak{B})$ for which $\xi(\alpha)[\eta(\beta)]$ is equal to 1.

Theorem 2.3. If one of the spaces A and B is conditionally compact, for any $\epsilon > 0$, there exists a finite subset α of A and a finite subset β of B such that the value of the game relative to (A, B) differs at most by ϵ from the value of each of the following three games: the game relative to (α, B), that relative to (A, β), and that relative to (α, β).

Proof: Let the subsets $A_1, \cdots, A_k$ of A and the subsets $B_1, \cdots, B_l$ of B be chosen as in the proof of Theorem 2.2. Also let $\alpha = \{a_1, \cdots, a_k\}$ and $\beta = \{b_1, \cdots, b_l\}$, where a_i is an element in A_i and b_j is an element in B_j. It follows from (2.26) and Theorem 2.2 that the value of the game relative to (α, B) differs from that of the game relative to (A, B) at most by ϵ. Similarly one sees that the value of the game

relative to (A, β) differs from that corresponding to (A, B) at most by ϵ. Equation (2.28) and Theorem 2.2 imply that this is true also for the game relative to (α, β). Thus our theorem is proved.

Theorem 2.4. If one of the spaces, say the space A, is finite, then for any $\epsilon > 0$ there exists a finite subset β of B such that the number of points contained in β does not exceed the number of points contained in A and the value of the game is not changed by more than ϵ when B is replaced by β.

Proof: According to Theorem 2.3 there exists a finite subset β^* of B such that the value of the game is changed at most by ϵ when B is replaced by β^*. If β^* contains more points than A, then, according to a result by Kaplansky [27], we can replace β^* by a subset β of β^* such that the value of the game relative to (A, β) is the same as the value of the game relative to (A, β^*) and A and β contain the same number of elements. This proves our theorem.

Theorem 2.5. If one of the spaces, say A, consists of m points $(m < \infty)$ and if B is compact, then there exists a finite subset β of B such that β contains at most m points and the value of the game remains unchanged when B is replaced by β.

Proof: Let $\{\epsilon_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ be a sequence of positive numbers such that

$$\lim_{i=\infty} \epsilon_i = 0$$

According to Theorem 2.4 there exists a subset β_i of B such that β_i contains at most m points and the value of the game is changed at most by ϵ_i when B is replaced by β_i. Clearly there exists a subsequence $\{i_j\}$ $(j = 1, 2, \cdots, \text{ad inf.})$ of the sequence $\{i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ such that the number of points contained in β_{i_j} $(j = 1, 2, \cdots, \text{ad inf.})$ is equal to a fixed integer n, independent of j, and the points in β_{i_j} converge to some limit points as $j \to \infty$. Let β be the limit of β_{i_j} as $j \to \infty$. Since the value of the game corresponding to (A, β_{i_j}) converges to the value of the game corresponding to (A, β), the value of the game corresponding to (A, β) is equal to the value of the game corresponding to (A, B). Thus our theorem is proved.

2.1.4 The Case when the Space of Strategies of One of the Players Is Separable

A space C is said to be separable if there exists a countable subset γ of C that is dense in C, i.e., a subset γ with the property that for any point c in C there exists a sequence $\{c_i\}$ of points in γ such that

$$\lim_{i=\infty} c_i = c$$

Separability of one of the spaces A and B does not necessarily imply the separability of the other space, as shown by the following example. Let A be the space of all positive integers and B the space of all subsequences of the sequence of positive integers. Thus any element a of A is a positive integer, and any element b of B is a subsequence of the sequence of all positive integers. Let $K(a, b) = -1$ if a is not an element of the sequence b, and let $K(a, b) = 1$ when a is an element of b. In this case A is separable but B is not, since the distance between two different elements of B is always 2 and the number of elements in B is non-denumerable.

Theorem 2.6. If one of the spaces A and B, say A, is separable and if α is a dense subset of A, then the class of all probability measures ξ_α; i.e., the class of all probability measures ξ for which $\xi(\alpha) = 1$ is dense in the class of all probability measures ξ in the sense of the metric given in (2.4).

Proof: Let $\{\epsilon_i\}$ $(i = 1, 2, \cdots,$ ad inf.) be a sequence of positive numbers such that $\lim_{i=\infty} \epsilon_i = 0$. Since A is separable, there exists a sequence $\{\alpha_{i_1 \cdots i_k}\}$ $(i_1 = 1, 2, \cdots,$ ad inf., $i_2 = 1, 2, \cdots,$ ad inf., $\cdots$, $i_k = 1, 2, \cdots,$ ad inf., $k = 1, 2, \cdots,$ ad inf.) of subsets of A such that the following conditions are satisfied: (1) The sets $\alpha_{i_1 \cdots i_{kj}}$ and $\alpha_{i_1 \cdots i_{kj^*}}$ are disjoint for $j \neq j^*$; (2) $\sum_{i_k=1}^{\infty} \alpha_{i_1 \cdots i_k} = \alpha_{i_1 \cdots i_{k-1}}$; (3) $\sum_{i_1=1}^{\infty} \alpha_{i_1} = A$; (4) the diameter of $\alpha_{i_1 \cdots i_k}$ does not exceed ϵ_k; (5) the intersection $\bar{\alpha}_{i_1 \cdots i_k}$ of α and $\alpha_{i_1 \cdots i_k}$ is not empty. For any k, $i_1, \cdots, i_k$, let $a_{i_1 \cdots i_k}$ be a given point in $\bar{\alpha}_{i_1 \cdots i_k}$. For any probability measure ξ^0 and for any k, let $\xi_k{}^0$ be the probability measure for which $\xi_k{}^0(a_{i_1 \cdots i_k}) = \xi^0(\alpha_{i_1 \cdots i_k})$ for all values of $i_1, \cdots, i_k$. Clearly

$$|K(\xi^0, \eta) - K(\xi_k{}^0, \eta)| \leqq \epsilon_k \tag{2.31}$$

for all η. Hence

$$\lim_{k=\infty} K(\xi_k{}^0, \eta) = K(\xi^0, \eta) \tag{2.32}$$

uniformly in η. Theorem 2.6 is an immediate consequence of (2.32).

If A is separable, there exists a denumerable subset α that is dense in A. It then follows from Theorem 2.6 that the class of discrete probability measures ξ lies dense in the class of all probability measures ξ. A probability measure ξ is said to be discrete if there exists a denumerable subset α of A such that $\xi(\alpha) = 1$. Thus, if A is separable and if the game is strictly determined, the value of the game is not affected by the choice of the Borel field $\mathfrak{A}$, provided that $\mathfrak{A}$ contains all

denumerable subsets of A as elements. Hence, if A is separable, from the point of view of the value of the game it makes no difference what Borel field $\mathfrak{A}$ is adopted as long as $\mathfrak{A}$ contains all denumerable subsets of A as elements and $K(a, b)$ is measurable $(\mathfrak{C})$, where $\mathfrak{C}$ is the Borel field defined at the end of Section 2.1.1. The choice of $\mathfrak{A}$ as the smallest Borel field $\mathfrak{A}_1$ containing all open subsets of A seems to be satisfactory in every respect when A is separable, particularly since then $K(a, b)$ is always measurable $(\mathfrak{C})$ (provided that $\mathfrak{B} = \mathfrak{B}_1$). Thus there is little interest in considering Borel fields $\mathfrak{A}$ different from $\mathfrak{A}_1$.

Theorem 2.7. Let A be separable. Also let $\{\alpha_i\}$ be a sequence of subsets of A such that α_i is conditionally compact, $\alpha_i \subset \alpha_{i+1}$ $(i = 1, 2, \cdots, ad\ inf.)$, and $\sum_{i=1}^{\infty} \alpha_i = \alpha$ is dense in A. Then a necessary and sufficient condition for strict determinateness of the game is that

$$\lim_{i=\infty} \operatorname{Inf}_\eta \operatorname{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) = \inf_\eta \operatorname{Sup}_\xi K(\xi, \eta) \tag{2.33}$$

Proof: According to Lemma 2.1 we have

$$\lim_{i=\infty} \operatorname{Sup}_{\xi_{\alpha_i}} \operatorname{Inf}_\eta K(\xi_{\alpha_i}, \eta) = \operatorname{Sup}_{\xi_\alpha} \operatorname{Inf}_\eta K(\xi_\alpha, \eta) \tag{2.34}$$

Since α_i is conditionally compact, the game relative to (α_i, B) is strictly determined; i.e.,

$$\operatorname{Sup}_{\xi_{\alpha_i}} \operatorname{Inf}_\eta K(\xi_{\alpha_i}, \eta) = \operatorname{Inf}_\eta \operatorname{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) \tag{2.35}$$

From (2.34) and (2.35) we obtain

$$\lim_{i=\infty} \operatorname{Inf}_\eta \operatorname{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) = \operatorname{Sup}_{\xi_\alpha} \operatorname{Inf}_\eta K(\xi_\alpha, \eta) \tag{2.36}$$

From Theorem 2.6 it follows that

$$\operatorname{Sup}_{\xi_\alpha} \operatorname{Inf}_\eta K(\xi_\alpha, \eta) = \operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) \tag{2.37}$$

The equivalence of (2.33) with the strict determinateness of the game follows immediately from (2.36) and (2.37).

When the roles of the two players are interchanged, Theorem 2.7 immediately gives the following theorem.

Theorem 2.8. Let B be separable and let $\{\beta_i\}$ be a sequence of subsets of B such that β_i is conditionally compact, $\beta_i \subset \beta_{i+1}$, and $\sum_{i=1}^{\infty} \beta_i = \beta$ is dense in B. Then a necessary and sufficient condition for strict determinateness of the game is that

$$\lim_{i=\infty} \operatorname{Sup}_\xi \operatorname{Inf}_{\eta_{\beta_i}} K(\xi, \eta_{\beta_i}) = \operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) \tag{2.38}$$

Theorem 2.9. If one of the spaces A and B, say A, is separable and if the game is strictly determined, then

(i) *there exists a denumerable subset α of A such that the game relative to (α, B) is strictly determined and its value is equal to the value of the game relative to (A, B);*

(ii) *for any $\epsilon > 0$, there exists a finite subset α_ϵ of A such that the value of the game relative to (α_ϵ, B) differs from the value of the game relative to (A, B) at most by ϵ.*

Proof: Assume that A is separable and that the game is strictly determined. Because of the separability of A, there exists a sequence $\alpha = \{a_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of elements of A that is dense in A. It follows from Theorem 2.6 that the value of the game relative to (α, B) is the same as that of the game relative to (A, B), and statement (i) of our theorem is proved. Let α_i be the set consisting of the first i elements of the sequence $\{a_j\}$ $(j = 1, 2, \cdots,$ ad inf.). Since the game is strictly determined, it follows from Theorem 2.7 that

$$\lim_{i=\infty} \operatorname{Inf}_\eta \operatorname{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) = \operatorname{Inf}_\eta \operatorname{Sup}_\xi K(\xi, \eta) = \operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) \tag{2.39}$$

For any subset α^* of A and for any subset β^* of B we shall denote by $v(\alpha^*, \beta^*)$ the value of the game when A is replaced by α^* and B is replaced by β^*, provided that the game relative to (α^*, β^*) is strictly determined. Since the game relative to (α_i, B) is strictly determined, it follows from (2.39) that

$$\lim_{i=\infty} v(\alpha_i, B) = v(A, B) \tag{2.40}$$

Thus, for any $\epsilon > 0$, there exists a positive integer i such that

$$\left| v(\alpha_i, B) - v(A, B) \right| \leqq \epsilon \tag{2.41}$$

This completes the proof of Theorem 2.9.

Theorem 2.10. If both spaces A and B are separable and if the game is strictly determined, then

(i) *there exists a denumerable subset α of A and a denumerable subset β of B such that $v(A, B) = v(\alpha, B) = v(A, \beta) = v(\alpha, \beta)$;*

(ii) *for any $\epsilon > 0$, there exists a finite subset α_ϵ of A and a finite subset β_ϵ of B such that each of the values $v(\alpha_\epsilon, B)$, $v(A, \beta_\epsilon)$, and $v(\alpha_\epsilon, \beta_\epsilon)$ differs from $v(A, B)$ at most by ϵ.*

Proof: It follows from Theorem 2.9 that there exists a denumerable subset α of A and a denumerable subset β of B such that the games rela-

tive to (α, B) and (A, β) are strictly determined and

$$v(\alpha, B) = v(A, \beta) = v(A, B) \tag{2.42}$$

For any subset α^* of A and any subset β^* of B the following inequalities obviously hold.

$$\begin{aligned} \operatorname{Sup}_{\xi_{\alpha^*}} \operatorname{Inf}_{\eta} K(\xi_{\alpha^*}, \eta) &\leqq \operatorname{Sup}_{\xi_{\alpha^*}} \operatorname{Inf}_{\eta_{\beta^*}} K(\xi_{\alpha^*}, \eta_{\beta^*}) \\ &\leqq \operatorname{Sup}_{\xi} \operatorname{Inf}_{\eta_{\beta^*}} K(\xi, \eta_{\beta^*}) \end{aligned} \tag{2.43}$$

and

$$\begin{aligned} \operatorname{Inf}_{\eta} \operatorname{Sup}_{\xi_{\alpha^*}} K(\xi_{\alpha^*}, \eta) &\leqq \operatorname{Inf}_{\eta_{\beta^*}} \operatorname{Sup}_{\xi_{\alpha^*}} K(\xi_{\alpha^*}, \eta_{\beta^*}) \\ &\leqq \operatorname{Inf}_{\eta_{\beta^*}} \operatorname{Sup}_{\xi} K(\xi, \eta_{\beta^*}) \end{aligned} \tag{2.44}$$

Replacing α^* and β^* in (2.43) and (2.44) by α and β, respectively, it follows from (2.42)–(2.44) that the game relative to (α, β) is strictly determined and $v(\alpha, \beta) = v(A, B)$. Thus statement (i) is proved.

Let α_ϵ be a finite subset of A and β_ϵ a finite subset of B such that

$$\left| v(\alpha_\epsilon, B) - v(A, B) \right| \leqq \epsilon \qquad \text{and} \qquad \left| v(A, \beta_\epsilon) - v(A, B) \right| \leqq \epsilon \tag{2.45}$$

The existence of such subsets follows from Theorem 2.9. Replacing α^* and β^* in (2.43) and (2.44) by α_ϵ and β_ϵ, respectively, it follows from (2.43) that

$$v(\alpha_\epsilon, B) \leqq v(\alpha_\epsilon, \beta_\epsilon) \leqq v(A, \beta_\epsilon) \tag{2.46}$$

Statement (ii) is an immediate consequence of (2.45) and (2.46), and the proof of Theorem 2.10 is completed.

2.1.5 General Spaces of Strategies

We shall now consider the general case where the spaces of strategies may be non-separable. In the case of a separable space, we have seen that the class of all discrete probability measures is dense in the class of all probability measures. This is not necessarily true for non-separable spaces. We shall, however, study the problem of strict determinateness under the restriction that the mixed strategy of a player must be either a discrete probability measure or a limit, in the sense of the distance definitions (2.4) and (2.5), of a sequence of discrete probability measures. This restriction is perhaps not too serious from the point of view of applications.

Theorem 2.11. If the mixed strategy used by player 1 must be either a discrete probability measure or a limit, in the sense of the distance definition (2.4), of a sequence of discrete probability measures, then a necessary and sufficient condition for the game to be strictly determined is that there

exists a sequence $\alpha = \{a_i\}$ $(i = 1, 2, \cdots,$ *ad inf.) of elements of* A *such that*

$$\lim_{i=\infty} \mathrm{Inf}_\eta \, \mathrm{Sup}_{\xi\alpha_i} K(\xi_{\alpha_i}, \eta) = \mathrm{Inf}_\eta \, \mathrm{Sup}_\xi K(\xi, \eta) \tag{2.47}$$

where $\alpha_i = \{a_1, \cdots, a_i\}$.

Proof: Because of the restriction imposed on the mixed strategy that can be used by player 1, there exists a sequence $\{\xi_i\}$ of discrete probability measures such that

$$\lim_{i=\infty} \mathrm{Inf}_\eta K(\xi_i, \eta) = \mathrm{Sup}_\xi \mathrm{Inf}_\eta K(\xi, \eta) \tag{2.48}$$

Since ξ_i is a discrete probability measure, there exists a countable subset $\bar{\alpha}_i$ such that $\xi_i(\bar{\alpha}_i) = 1$. Let $\alpha = \sum_{i=1}^{\infty} \bar{\alpha}_i$. It then follows from (2.48) that

$$\mathrm{Sup}_{\xi\alpha} \mathrm{Inf}_\eta K(\xi_\alpha, \eta) = \mathrm{Sup}_\xi \mathrm{Inf}_\eta K(\xi, \eta) \tag{2.49}$$

We arrange the elements of α in an ordered sequence. Let $\alpha = \{a_i\}$ $(i = 1, 2, \cdots,$ ad inf.). Since $\alpha_i = \{a_1, \cdots, a_i\}$ is finite, the game relative to (α_i, B) is strictly determined; i.e.,

$$\mathrm{Inf}_\eta \, \mathrm{Sup}_{\xi\alpha_i} K(\xi_{\alpha_i}, \eta) = \mathrm{Sup}_{\xi\alpha_i} \mathrm{Inf}_\eta K(\xi_{\alpha_i}, \eta) \tag{2.50}$$

It then follows from Lemma 2.1 that

$$\begin{aligned} \lim_{i=\infty} \mathrm{Inf}_\eta \, \mathrm{Sup}_{\xi\alpha_i} K(\xi_{\alpha_i}, \eta) &= \mathrm{Sup}_{\xi\alpha} \mathrm{Inf}_\eta K(\xi_\alpha, \eta) \\ &= \mathrm{Sup}_\xi \mathrm{Inf}_\eta K(\xi, \eta) \end{aligned} \tag{2.51}$$

The necessity of our condition follows immediately from (2.51). To prove sufficiency, let $\alpha = \{a_i\}$ be a sequence satisfying (2.47). Clearly (2.50) and the first half of (2.51) are satisfied for this sequence. It then follows from Lemma 2.4 that the game is strictly determined and the sufficiency of our condition is proved.

When the roles of the two players are interchanged, Theorem 2.11 gives the following theorem.

Theorem 2.12. If the mixed strategy used by player 2 must be either a discrete probability measure or a limit, in the sense of the metric (2.5), of a sequence of discrete probability measures, then a necessary and sufficient condition for the game to be strictly determined is that there exists a sequence $\beta = \{b_i\}$ *of elements of* B *such that*

$$\lim_{i=\infty} \mathrm{Sup}_\xi \mathrm{Inf}_{\eta_{\beta_i}} K(\xi, \eta_{\beta_i}) = \mathrm{Sup}_\xi \mathrm{Inf}_\eta K(\xi, \eta) \tag{2.52}$$

where $\beta_i = \{b_1, \cdots, b_i\}$.

We shall now prove the following theorem.

Theorem 2.13. If the space of mixed strategies of one of the players, say player 1, is restricted to the closure, in the sense of the metric (2.4), of all discrete probability measures ξ and if the game is strictly determined, then there exists a countable subset α of A such that $v(\alpha, B) = v(A, B)$, and for any $\epsilon > 0$ there exists a finite subset α_ϵ of A such that $| v(\alpha_\epsilon, B) - v(A, B) | \leqq \epsilon$. If the space of mixed strategies is restricted for each of the players to the closure of all discrete probability measures and if the game is strictly determined, there exists a countable subset α of A and a countable subset β of B such that $v(\alpha, B) = v(A, \beta) = v(\alpha, \beta) = v(A, B)$, and for any $\epsilon > 0$ there exists a finite subset α_ϵ of A and a finite subset β_ϵ of B such that each of the values $v(\alpha_\epsilon, B)$, $v(A, \beta_\epsilon)$, and $v(\alpha_\epsilon, \beta_\epsilon)$ differs from $v(A, B)$ at most by ϵ.

Proof: Suppose that the space of mixed strategies of player 1 is restricted to the closure of all discrete probability measures and that the game is strictly determined. Then according to Theorem 2.11 there exists a sequence $\alpha = \{a_i\}$ of elements of A for which (2.47) holds. Thus

$$\lim_{i=\infty} v(\alpha_i, B) = v(A, B) \tag{2.53}$$

Theorem 2.7 and (2.47) imply that the game relative to (α, B) is strictly determined and that

$$\lim_{i=\infty} v(\alpha_i, B) = v(\alpha, B) \tag{2.54}$$

The first part of Theorem 2.13 is an immediate consequence of (2.53) and (2.54). To prove the second half of our theorem, assume that the space of mixed strategies of each of the players is restricted to the closure of the class of all discrete probability measures and that the game is strictly determined. Let α be a countable subset of A, α_ϵ a finite subset of A, β a countable subset of B, and β_ϵ a finite subset of B such that

$$v(\alpha, B) = v(A, \beta) = v(A, B) \tag{2.55}$$

and

$$| v(\alpha_\epsilon, B) - v(A, B) | \leqq \epsilon, \quad | v(A, \beta_\epsilon) - v(A, B) | \leqq \epsilon \tag{2.56}$$

The existence of such subsets follows from the first half of our theorem.

It follows from (2.43), (2.44), and (2.55) that the game relative to (α, β) is strictly determined and that $v(\alpha, \beta) = v(A, B)$. Furthermore

(2.43) and (2.44) imply that

$$v(\alpha_\epsilon, B) \leqq v(\alpha_\epsilon, \beta_\epsilon) \leqq v(A, \beta_\epsilon) \tag{2.57}$$

Hence

$$\left| v(\alpha_\epsilon, \beta_\epsilon) - v(A, B) \right| \leqq \epsilon \tag{2.58}$$

This completes the proof of our theorem.

If the restrictions imposed on the mixed strategies that can be used by the players are lifted, the conclusions in Theorems 2.11, 2.12, and 2.13 do not necessarily hold, as shown by the following example. Let the space A, as well as the space B, be the open interval (0, 1) on the real line. Then any element a of A and any element b of B can be written in dyadic form as a sequence of 0's and 1's.[7] For each positive integer k let S_k be a subsequence of the sequence S of all positive integers such that $\sum_{k=1}^{\infty} S_k = S$ and $S_1, S_2, \cdots$, etc., are disjoint. Let $K(a, b) = -1$ if there exists a positive integer k such that $a_i = b_i$ for all i in S_k, where the sequence $\{a_i\}$ $(i = 1, 2, \cdots,$ ad inf.) is the dyadic representation of a and the sequence $\{b_i\}$ is the dyadic representation of b. In all other cases we put $K(a, b) = 1$. Let ξ^0 be the uniform distribution on the interval (0,1); i.e., $\xi^0(\alpha)$ is equal to the Lebesgue measure of α. Clearly

$$K(\xi^0, b) = 1 \tag{2.59}$$

for all b. Hence

$$K(\xi^0, \eta) = 1 \tag{2.60}$$

for all η. Hence, if $\mathfrak{A}$ is the class of all Lebesgue measurable subsets of the interval (0, 1), we have

$$\operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) = \operatorname{Inf}_\eta \operatorname{Sup}_\xi K(\xi, \eta) = 1 \tag{2.61}$$

Now let $\alpha = \{a_i\}$ $(i = 1, 2, \cdots,$ ad inf.) be a sequence of elements of A and let $\{a_{ij}\}$ $(j = 1, 2, \cdots,$ ad inf.) be the dyadic representation of a_i. Let b_0 be the element of B whose dyadic representation $\{b_{0j}\}$ $(j = 1, 2, \cdots,$ ad inf.) is given as follows: $b_{0r} = a_{kr}$ for all r in S_k $(k = 1, 2, \cdots,$ ad inf.). Clearly

$$K(\xi_\alpha, b_0) = -1$$

identically in ξ_α.

[7] To make the dyadic representation unique for the dyadic rational points of the open interval (0, 1), we shall agree to take the representation which contains infinitely many zeros.

Hence for any denumerable subset α we have

$$\text{(2.62)} \qquad \operatorname{Sup}_{\xi_\alpha} \operatorname{Inf}_\eta K(\xi_\alpha, \eta) = \operatorname{Inf}_\eta \operatorname{Sup}_{\xi_\alpha} K(\xi_\alpha, \eta) = -1$$

This contradicts the conclusions in Theorems 2.11, 2.12, and 2.13.

It follows from (2.62) that, if ξ is restricted to discrete probability measures, $\operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) = -1$. On the other hand, $\operatorname{Inf}_\eta \operatorname{Sup}_\xi K(\xi, \eta)$ remains equal to 1 even when ξ is restricted to discrete probability measures.[8] Thus, if player 1 is restricted to the choice of discrete probability measures ξ, the game is not strictly determined. If ξ is not restricted to discrete probability measures, the choice of $\mathfrak{A}$ as the smallest Borel field containing all open subsets of A would not be satisfactory here, since any subset of A is open and thus $\mathfrak{A}$ would be the class of all subsets of A, which would narrow down the class of possible probability measures ξ unnecessarily in view of the (total) additivity condition imposed on ξ. The reason why any subset of A is open is that the distance between any two different elements a_1 and a_2 of A is equal to 2. This can be seen as follows: Let $\{a_{1j}\}$ and $\{a_{2j}\}$ ($j = 1, 2, \cdots$, ad inf.) be the dyadic representations of a_1 and a_2, respectively. Let b_0 be an element of B whose dyadic representation $\{b_{0j}\}$ ($j = 1, 2, \cdots$, ad inf.) satisfies the following conditions: Let r be the smallest positive integer with the property that $a_{1j} \neq a_{2j}$ for at least one value j in S_r. We put $b_{0j} = a_{1j}$ for all j in S_r. For any $k \neq r$, we put $b_{0j_k} = 1 - a_{2j_k}$, where j_k is the smallest integer in S_k. For all other integers j, the value b_{0j} may be determined arbitrarily subject to the only condition that the sequence $\{b_{0j}\}$ should contain infinitely many zeros. Then we have $K(a_1, b_0) = -1$ and $K(a_2, b_0) = 1$. Hence $\delta(a_1, a_2) = 2$.

2.2 Theorems Concerning the Topology of the Spaces of Mixed Strategies

2.2.1 Two Convergence Definitions in the Spaces of Mixed Strategies and Their Relations

An intrinsic metric in the spaces of mixed strategies has been introduced in Section 2.1.1 [see equations (2.4) and (2.5)]. We shall say that a sequence $\{\xi_i\}$ of probability measures in A converges in the intrinsic sense to the probability measure ξ_0 if $\lim_{i=\infty} \delta(\xi_i, \xi_0) = 0$, where $\delta(\xi_i, \xi_0)$ denotes the distance defined in (2.4). Intrinsic convergence of probability measures in the space B is defined similarly.

Another definition of convergence in the spaces of mixed strategies,

[8] One can easily verify that for any η there exists an element a in A such that $K(a, \eta) = 1$.

more in accordance with the ordinary notion of convergence, is this: We shall say that a sequence $\{\xi_i\}$ of probability measures in A converges in the ordinary sense to the probability measure ξ_0 if for any open subset α of A whose boundary has the probability 0 according to ξ_0 we have $\lim_{i=\infty} \xi_i(\alpha) = \xi_0(\alpha)$. Ordinary convergence in the space of mixed strategies of player 2 is defined similarly.

Theorem 2.14. If the space A is separable and if $\{\xi_i\}$ ($i = 1, 2, \cdots$, ad inf.) is a sequence of probability measures in A such that ξ_i converges to ξ_0 in the ordinary sense, then ξ_i converges to ξ_0 in the intrinsic sense also.

Proof: Let $\{\xi_i\}$ be a sequence of probability measures in A such that $\lim_{i=\infty} \xi_i = \xi_0$ in the ordinary sense. For any $\delta > 0$, there exists a sequence $\{\alpha_i\}$ ($i = 1, 2, \cdots$, ad inf.) of disjoint and non-empty open subsets of A such that for any i the diameter of α_i does not exceed δ, $\sum_{i=1}^{\infty} \bar{\alpha}_i = A$, and $\xi_0(\bar{\alpha}_i - \alpha_i) = 0$ ($i = 1, 2, \cdots$, ad inf.). Here $\bar{\alpha}_i$ denotes the closure of α_i. Let a_i be a particular point of α_i and ξ_n^* ($n = 0, 1, 2, \cdots$, ad inf.) be the probability measure that assigns to a_i the same probability as ξ_n to α_i; i.e., $\xi_n^*(a_i) = \xi_n(\alpha_i)$ ($i = 1, 2, \cdots$, ad inf.). Clearly $\lim_{n=\infty} \xi_n^* = \xi_0^*$ in the ordinary sense and, because of the uniform boundedness of $K(a, b)$, we have

$$\lim_{n=\infty} K(\xi_n^*, \eta) = K(\xi_0^*, \eta) \tag{2.63}$$

uniformly in η. Since the diameter of α_i does not exceed δ, we have

$$\left| K(\xi_i^*, \eta) - K(\xi_i, \eta) \right| \leqq \delta \tag{2.64}$$

for all η and for $i = 0, 1, 2, \cdots$, ad inf. Since δ can be chosen arbitrarily small, it follows from (2.63) and (2.64) that

$$\lim_{n=\infty} K(\xi_n, \eta) = K(\xi_0, \eta) \tag{2.65}$$

uniformly in η. But (2.65) is equivalent to convergence of ξ_n to ξ_0 in the intrinsic sense, and Theorem 2.14 is proved.

2.2.2 Compactness of the Space of Mixed Strategies when the Space of Pure Strategies Is Compact

We shall show in this section that, if the space A of pure strategies is compact, the space of all mixed strategies ξ is also compact in the sense of intrinsic as well as ordinary convergence. It is sufficient to prove compactness of the space of all mixed strategies in the sense of ordinary convergence, since, according to Theorem 2.14, this also

implies compactness in the intrinsic sense. More precisely, we shall prove the following theorem.

Theorem 2.15. If A is compact and if $\{\xi_n\}$ is a sequence of probability measures in A, the sequence $\{\xi_n\}$ has a subsequence that converges in the ordinary sense to a limit probability measure.[9]

Proof: Assume that A is compact and let $\{\epsilon_k\}$ $(k = 1, 2, \cdots, \text{ad inf.})$ be a sequence of positive numbers such that $\lim_{k=\infty} \epsilon_k = 0$. Let $A_1, A_2, \cdots, A_{m_1}$ be mutually disjoint open subsets of A such that $\bar{A}_1 + \cdots + \bar{A}_{m_1} = A$ and the diameter of $\bar{A}_i$ does not exceed ϵ_1.[10] In general let $\{A_{i_1 i_2 \cdots i_k}\}$ $(i_j = 1, \cdots, m_j; j = 1, \cdots, k)$ be a system of $m_1 m_2 \cdots m_k$ open and disjoint subsets of A such that

$$\sum_{i_k=1}^{m_k} \bar{A}_{i_1 i_2 \cdots i_k} = \bar{A}_{i_1 i_2 \cdots i_{k-1}} \quad (k = 2, 3, \cdots, \text{ad inf.}) \tag{2.66}$$

and the diameter of $A_{i_1 i_2 \cdots i_k}$ does not exceed ϵ_k. Moreover the sets $A_{i_1 \cdots i_k}$ are chosen so that

$$\xi_n(\bar{A}_{i_1 \cdots i_k} - A_{i_1 \cdots i_k}) = 0 \tag{2.67}$$

for all values of $n, k, i_1, \cdots, i_k$.

Using the well-known diagonal procedure, we can construct a subsequence $\{\xi_{n_j}\}$ $(j = 1, 2, \cdots, \text{ad inf.})$ of the sequence $\{\xi_n\}$ such that $\lim_{j=\infty} \xi_{n_j}(A_{i_1 \cdots i_k})$ exists for any $k, i_1, \cdots, i_k$. For any subset α of A, let $\xi^*(\alpha)$ denote the limit of $\xi_{n_j}(\alpha)$ as $j \to \infty$, provided that the limit exists. Thus we can write

$$\lim_{j=\infty} \xi_{n_j}(A_{i_1 \cdots i_k}) = \xi^*(A_{i_1 \cdots i_k}) \tag{2.68}$$

Since $\xi_n(\bar{A}_{i_1 \cdots i_k}) = \xi_n(A_{i_1 \cdots i_k})$, we have $\xi^*(\bar{A}_{i_1 \cdots i_k}) = \xi^*(A_{i_1 \cdots i_k})$. For any open subset α of A, let $\xi(\alpha)$ denote the least upper bound of $\xi^*(\bar{\beta})$ with respect to $\bar{\beta}$, where $\bar{\beta}$ may be the sum of any finite number of sets $\bar{A}_{i_1 \cdots i_k}$ that are contained in α. Clearly $\xi(\alpha) \geqq 0$, $\xi(A) = 1$, and $\xi(\alpha_1 + \alpha_2) = \xi(\alpha_1) + \xi(\alpha_2)$ for any open and disjoint sets α_1 and α_2. Let $\{\alpha_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ be a sequence of open and disjoint subsets of A. Let $\bar{\beta}$ be the sum of a finite number of sets $\bar{A}_{i_1 \cdots i_k}$ such that $\bar{\beta}$ is contained in $\alpha = \sum_{i=1}^{\infty} \alpha_i$. Since A is compact, it follows from

[9] A closely related theorem was proved by Kryloff and Bogoliouboff [28]. Their convergence definition in the space of the probability measure ξ is somewhat different from the one used here.

[10] In what follows in this section, for any subset α of A we shall use the symbol $\bar{\alpha}$ to denote the closure of α.

the Borel covering theorem that $\bar{\beta}$ will be contained in the sum of a finite number of elements of the sequence $\{\alpha_i\}$. From this it follows that $\xi(\sum_{i=1}^{\infty} \alpha_i) = \sum_{i=1}^{\infty} \xi(\alpha_i)$. The measure function $\xi(\alpha)$ can be extended in the usual way to all elements α of the smallest Borel field containing all open subsets of A.

Let α be an open subset of A such that

$$\xi(\bar{\alpha} - \alpha) = 0 \tag{2.69}$$

For any positive integer k, let $\bar{\beta}_k$ be the sum of all those sets $\bar{A}_{i_1 \dots i_k}$ which have common points with α but are not included in α. Since $\xi(\bar{\alpha} - \alpha) = 0$, we must have

$$\lim_{k=\infty} \xi^*(\bar{\beta}_k) = 0 \tag{2.70}$$

For any k, let $\bar{\gamma}_k$ be the sum of all those $\bar{A}_{i_1 \dots i_k}$ which are included in α. Clearly

$$\xi_{n_j}(\bar{\gamma}_k + \bar{\beta}_k) \geqq \xi_{n_j}(\alpha) \geqq \xi_{n_j}(\bar{\gamma}_k) \tag{2.71}$$

Hence

$$\xi^*(\bar{\gamma}_k + \bar{\beta}_k) \geqq \limsup_{j=\infty} \xi_{n_j}(\alpha) \geqq \liminf_{j=\infty} \xi_{n_j}(\alpha) \geqq \xi^*(\bar{\gamma}_k) \tag{2.72}$$

Since

$$\xi^*(\bar{\gamma}_k + \bar{\beta}_k) \geqq \xi(\alpha) \geqq \xi^*(\bar{\gamma}_k) \tag{2.73}$$

and since

$$\lim_{k=\infty} [\xi^*(\bar{\gamma}_k + \bar{\beta}_k) - \xi^*(\bar{\gamma}_k)] = 0 \tag{2.74}$$

it follows from (2.72) that

$$\lim_{j=\infty} \xi_{n_j}(\alpha) = \xi(\alpha) \tag{2.75}$$

Hence Theorem 2.15 is proved.

2.2.3 Separability of the Space of Mixed Strategies when the Space of Pure Strategies Is Separable

The purpose of this section is to prove the following theorem.

Theorem 2.16. If the space of pure strategies of player i is separable, the space of mixed strategies of player i is also separable in the sense of intrinsic convergence.

Proof: Let us assume that the space A of pure strategies of player 1 is separable. Let α_0 be a denumerable and dense subset of A. It

follows from Theorem 2.6 that the set of all probability measures ξ for which $\xi(\alpha_0) = 1$ is dense in the set of all ξ. The set of all probability measures ξ for which $\xi(\alpha_0) = 1$ is obviously separable in the sense of the ordinary convergence definition. Thus, because of Theorem 2.14, the set of all ξ for which $\xi(\alpha_0) = 1$ is separable also in the sense of the intrinsic convergence definition. Hence the space of all ξ must be separable in the sense of the intrinsic convergence definition, and Theorem 2.16 is proved.

2.3 Properties of Minimax Strategies [11]

In this section we shall state and prove some theorems concerning minimax strategies.

Theorem 2.17. If ξ_0 is a minimax strategy of player 1 and if the game is strictly determined, ξ_0 is a maximal strategy in the wide sense. Similarly, if η_0 is a minimax strategy of player 2 and the game is strictly determined, η_0 is a minimal strategy in the wide sense.

Proof: Suppose that the game is strictly determined and ξ_0 is a minimax strategy of player 1. Let $\{\eta_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ be a sequence of strategies of player 2 such that

$$\lim_{i=\infty} \operatorname{Sup}_\xi K(\xi, \eta_i) = \operatorname{Inf}_\eta \operatorname{Sup}_\xi K(\xi, \eta) \tag{2.76}$$

Since ξ_0 is a minimax strategy, we have

$$\operatorname{Inf}_\eta K(\xi_0, \eta) = \operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) \tag{2.77}$$

Hence

$$\operatorname{Sup}_\xi K(\xi, \eta_i) \geqq K(\xi_0, \eta_i) \geqq \operatorname{Sup}_\xi \operatorname{Inf}_\eta K(\xi, \eta) \tag{2.78}$$

Since the game is strictly determined, it follows from (2.76) and (2.78) that

$$\lim_{i=\infty} [\operatorname{Sup}_\xi K(\xi, \eta_i) - K(\xi_0, \eta_i)] = 0 \tag{2.79}$$

Hence ξ_0 is a maximal strategy in the wide sense. The second half of our theorem is proved by interchanging the roles of the two players.

Theorem 2.18. If the game is strictly determined, and if ξ_0 and η_0 are minimax strategies of players 1 and 2, respectively, then ξ_0 is a maximal strategy relative to η_0, η_0 is a minimal strategy relative to ξ_0, and

$$K(\xi_0, \eta_0) = \operatorname{Min}_\eta \operatorname{Max}_\xi K(\xi, \eta) = \operatorname{Max}_\xi \operatorname{Min}_\eta K(\xi, \eta) \tag{2.80}$$

[11] Most of the results of this section have been stated and proved by von Neumann for finite spaces of strategies. See Sections 17.8 and 17.9 of [55].

Proof: Since ξ_0 and η_0 are minimax strategies, the following inequalities must hold:

$$\text{(2.81)}\quad \operatorname{Max}_\xi \operatorname{Min}_\eta K(\xi, \eta) = \operatorname{Min}_\eta K(\xi_0, \eta) \leqq K(\xi_0, \eta_0) \leqq \operatorname{Max}_\xi K(\xi, \eta_0) = \operatorname{Min}_\eta \operatorname{Max}_\xi K(\xi, \eta)$$

Since the game is strictly determined, it follows from (2.81) that

$$\text{(2.82)}\qquad \operatorname{Min}_\eta K(\xi_0, \eta) = K(\xi_0, \eta_0) = \operatorname{Max}_\xi K(\xi, \eta_0)$$

and our theorem is proved.

Theorem 2.19. If the game is strictly determined and if ξ_0 is a minimax strategy of player 1 and η_0 is a minimax strategy of player 2, then

$$\text{(2.83)}\qquad \xi_0(A - \alpha_0) = \eta_0(B - \beta_0) = 0$$

where α_0 is the set of all elements a_0 of A for which $K(a_0, \eta_0) = Max_a K(a, \eta_0)$, and β_0 is the set of all elements b_0 of B for which $K(\xi_0, b_0) = Min_b K(\xi_0, b)$.

Proof: Clearly, for any probability measure ξ for which $\xi(A - \alpha_0) > 0$, we must have $K(\xi, \eta_0) < \operatorname{Max}_a K(a, \eta_0) = K(\xi_0, \eta_0)$. Hence $\xi_0(A - \alpha_0) = 0$. Similarly we can see that $\eta_0(B - \beta_0) = 0$.

We shall say that the space B is weakly compact [12] if for any sequence $\{\eta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of probability measures in B there exist a subsequence $\{\eta_{i_j}\}$ $(j = 1, 2, \cdots,$ ad inf.) and a probability measure η_0 such that

$$\text{(2.84)}\qquad \liminf_{j=\infty} K(\xi, \eta_{i_j}) \geqq K(\xi, \eta_0)$$

for all ξ. This is a weaker property than compactness. Even the existence of a subsequence $\{\eta_{i_j}\}$ and that of a probability measure η_0 such that $\lim_{j=\infty} K(\xi, \eta_{i_j}) = K(\xi, \eta_0)$ for all ξ is weaker than compactness, since compactness requires that the above convergence be uniform in ξ. The case when the space B is weakly compact in the sense of the above definition will play an important role in the theory of statistical decision functions. It will be seen in Chapter 3 that the space of strategies of the statistician is weakly compact under very general conditions.

Theorem 2.20. If the space B is weakly compact, a minimax strategy for player 2 exists.

[12] The term "weak compactness" used here has no relation to the same term used in the literature with reference to a set of functions. See, for example, Widder [72].

Proof: Let $\{\eta_i\}$ be a sequence of probability measures in B such that

$$\lim_{i=\infty} \operatorname{Sup}_\xi K(\xi, \eta_i) = \operatorname{Inf}_\eta \operatorname{Sup}_\xi K(\xi, \eta) \tag{2.85}$$

A sequence $\{\eta_i\}$ with the above property evidently exists. Since B is weakly compact, there exist a subsequence $\{\eta_{i_j}\}$ $(j = 1, 2, \cdots,$ ad inf.) of the sequence $\{\eta_i\}$ and a probability measure η_0 such that

$$\liminf_{j=\infty} K(\xi, \eta_{i_j}) \geqq K(\xi, \eta_0) \tag{2.86}$$

for all ξ. From (2.85) and (2.86) it follows that

$$\operatorname{Sup}_\xi K(\xi, \eta_0) = \operatorname{Inf}_\eta \operatorname{Sup}_\xi K(\xi, \eta)$$

Thus η_0 is a minimax strategy and Theorem 2.20 is proved.

2.4 Admissible Strategies and Complete Classes of Strategies

2.4.1 Minimal Complete Class of Strategies

In Section 1.6.2 the notions of admissible strategies and complete classes of strategies were defined. A complete class C of strategies will be said to be a minimal complete class if no proper subclass of C is a complete class.

Theorem 2.21. For each player there exists at most one minimal complete class of strategies. If a minimal complete class of strategies exists, it must be identical with the class C_0 of all admissible strategies.

The proof is omitted, since it is essentially the same as that given in Section 1.3 in connection with decision functions.

One can easily give examples of games where the class of all admissible strategies is not complete. As a matter of fact, one can easily construct games for which the class of all admissible strategies is empty. For example, let A consist of a single element a and B of a sequence $\{b_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of elements. Let $K(a, b_i) = 1/i$. Clearly in this case there is no admissible strategy for player 2.

Theorem 2.22. If A is separable and B is weakly compact, the class of all admissible strategies of player 2 is a complete class.[13]

[13] This theorem is related to a theorem of Zorn on partially ordered sets (see Zorn [76] and Lefschetz [29], page 5) but cannot be derived from it, since Zorn assumes that each simply ordered subset has an upper bound in the system, whereas in our case merely each denumerable and simply ordered subset can be shown to have an upper bound. Our theorem, however, could be derived (without the use of transfinite induction) from some more general results by Milgram [32] which contain Zorn's theorem as a special case.

Proof: Suppose that A is separable, B is weakly compact, and the class of all admissible strategies of player 2 is not complete. Then there exists a non-admissible strategy η_1 such that any strategy η that is uniformly better than η_1 is also non-admissible. Since η_1 is not admissible there exists at least one strategy η_2 that is uniformly better than η_1. Since η_2 itself is non-admissible, there must be a strategy η_3 that is uniformly better than η_2, and so on. In this way we obtain a sequence $\{\eta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of strategies such that η_j is uniformly better than η_i for $j > i$. Because of the weak compactness of B there exists a strategy $\eta_{\omega+1}$ that is uniformly better than any element of the sequence $\{\eta_i\}$. But then there exists a strategy $\eta_{\omega+2}$ that is uniformly better than $\eta_{\omega+1}$, and so on. Continuing this procedure, we obtain a non-denumerable well-ordered set S of strategies η such that any element of this well-ordered set is uniformly better than all the preceding ones.

Since A is separable, there exists a sequence $\{a_i\}$ of elements of A that is dense in A. Thus, if η' and η'' are two strategies such that η'' is uniformly better than η', there exists an integer i such that $K(a_i, \eta'') < K(a_i, \eta')$. For any positive integer i let S_i be the well-ordered subset of S given as follows: An element η of S belongs to S_i if and only if there exists an element η' in S that is an immediate predecessor of η and $K(a_i, \eta) < K(a_i, \eta')$. Clearly $\Sigma S_i = S'$, where S' is the set of all those elements of S which have immediate predecessors in S. Since S' is non-denumerable, there exists a positive integer i such that S_i is non-denumerable. Clearly, for any two elements η' and η'' of S_i such that η' precedes η'', we have $K(a_i, \eta') > K(a_i, \eta'')$. But this is impossible, and Theorem 2.22 is proved.

2.4.2 Theorems on Complete Classes of Strategies

In this section we shall derive several theorems concerning complete classes of strategies. First we shall prove the following two theorems.

Theorem 2.23. If A is separable and B is weakly compact, the game is strictly determined.

Proof: Since A is separable, there exists a sequence $\{\alpha_i\}$ of subsets of A such that $\alpha_i \subset \alpha_{i+1}$, α_i is conditionally compact, and

$$\sum_{i=1}^{\infty} \alpha_i = \alpha$$

is dense in A. The game relative to (α_i, B) is strictly determined. Let η_i be a minimax strategy for the game relative to (α_i, B). Because

of the weak compactness of B such a minimax strategy exists. Thus

$$\text{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta_i) = \text{Inf}_\eta \text{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) \tag{2.87}$$

Since B is weakly compact, there exist a subsequence $\{\eta_{i_j}\}$ $(j = 1, 2, \cdots, \text{ad inf.})$ of the sequence $\{\eta_i\}$ and a strategy η_0 such that

$$\liminf_{j=\infty} K(\xi, \eta_{i_j}) \geqq K(\xi, \eta_0) \tag{2.88}$$

for all ξ. From (2.87) and (2.88) it follows that for any positive integer r

$$\begin{aligned} \text{Sup}_{\xi_{\alpha_r}} K(\xi_{\alpha_r}, \eta_0) &\leqq \liminf_{j=\infty} \text{Sup}_{\xi_{\alpha_r}} K(\xi_{\alpha_r}, \eta_{i_j}) \\ &\leqq \lim_{j=\infty} \text{Sup}_{\xi_{\alpha_{i_j}}} K(\xi_{\alpha_{i_j}}, \eta_{i_j}) \\ &= \lim_{j=\infty} \text{Inf}_\eta \text{Sup}_{\xi_{\alpha_{i_j}}} K(\xi_{\alpha_{i_j}}, \eta) \\ &= \lim_{i=\infty} \text{Inf}_\eta \text{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) \end{aligned} \tag{2.89}$$

Hence, since $\lim_{r=\infty} \text{Sup}_{\xi_{\alpha_r}} K(\xi_{\alpha_r}, \eta_0) = \text{Sup}_{\xi_\alpha} K(\xi_\alpha, \eta_0)$,

$$\text{Inf}_\eta \text{Sup}_{\xi_\alpha} K(\xi_\alpha, \eta) \leqq \lim_{i=\infty} \text{Inf}_\eta \text{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta) \tag{2.90}$$

Obviously the left-hand member of (2.90) cannot be smaller than the right-hand member of (2.90), and therefore the equality sign must hold. Theorem 2.23 is an immediate consequence of this and Theorems 2.6 and 2.7.

Theorem 2.24. If A is separable, B is weakly compact, and the choice of player 2 is restricted to a class C of probability measures η which contains all discrete probability measures η and which does not destroy the property of weak compactness, then the game is strictly determined and its value is the same as if no restriction were imposed on the choice of η.

Proof: Let C be a class of probability measures η satisfying the conditions of our theorem. One can easily verify that equations (2.87) to (2.90) remain valid when η is restricted to elements of C. Hence the game remains strictly determined under the restriction that η must be an element of C. Let the sequence $\{\alpha_i\}$ be defined as in the proof of Theorem 2.23. It follows from Theorem 2.3 that the value of $\text{Inf}_\eta \text{Sup}_{\xi_{\alpha_i}} K(\xi_{\alpha_i}, \eta)$ remains unchanged when the choice of η is restricted to elements of C. This and (2.90) imply that the value of the game is not changed by restricting the choice of η to elements of C. Hence our theorem is proved.

Theorem 2.25. If A is separable and B is weakly compact, then for player 2 the class of minimal strategies in the wide sense is a complete class.

Proof: Let A be separable, B weakly compact, and η_0 a strategy of player 2 that is not a minimal strategy in the wide sense. We introduce a new outcome function $K^*(a, b)$ given by

$$K^*(a, b) = K(a, b) - K(a, \eta_0) \tag{2.91}$$

Clearly the space A remains separable, and the space B remains weakly compact, when $K(a, b)$ is replaced by $K^*(a, b)$. Thus the game corresponding to the outcome function $K^*(a, b)$ is strictly determined. Let η_1 be a minimax strategy for the game corresponding to $K^*(a, b)$. Because of the weak compactness of B such a minimax strategy exists. Since

$$K^*(\xi, \eta_0) = 0 \tag{2.92}$$

for all ξ, we must have

$$K^*(\xi, \eta_1) \leqq 0 \tag{2.93}$$

for all ξ. But

$$K^*(\xi, \eta) = K(\xi, \eta) - K(\xi, \eta_0) \tag{2.94}$$

Hence

$$K(\xi, \eta_1) \leqq K(\xi, \eta_0) \tag{2.95}$$

for all ξ. According to Theorem 2.17, η_1 is a minimal strategy in the wide sense when $K(a, b)$ is replaced by $K^*(a, b)$. But any strategy that is minimal in the wide sense relative to the outcome function $K^*(a, b)$ is also a minimal strategy in the wide sense relative to the original outcome function $K(a, b)$. Since η_0 is not a minimal strategy in the wide sense, the inequality sign must hold in (2.95) at least for some ξ. Thus η_1 is uniformly better than η_0, and our theorem is proved.

Theorem 2.26. If A is compact and B is weakly compact, then for player 2 the class of minimal strategies in the strict sense is a complete class.

Proof: Let A be compact, B be weakly compact, and η_0 be a strategy that is not a minimal strategy in the strict sense. Consider the outcome function $K^*(a, b) = K(a, b) - K(a, \eta_0)$. Clearly the space A remains compact and the space B remains weakly compact when $K(a, b)$ is replaced by $K^*(a, b)$. Let η_1 be a minimax strategy when $K^*(a, b)$ is the outcome function. Then, since $K^*(\xi, \eta_0) = 0$ identically in ξ, we have

$$K^*(\xi, \eta_1) = K(\xi, \eta_1) - K(\xi, \eta_0) \leqq 0 \tag{2.96}$$

identically in ξ. Let ξ_1 be a minimax strategy of player 1 when $K^*(a, b)$ is the outcome function. Such a minimax strategy exists, since A is compact. Then, according to Theorem 2.18, η_1 is a minimal strategy relative to ξ_1 when $K^*(a, b)$ is the outcome function. Clearly η_1 remains a minimal strategy relative to ξ_1 when $K(a, b)$ is the outcome function. Since η_0 is not a minimal strategy in the strict sense, $K(\xi, \eta_1) \neq K(\xi, \eta_0)$ at least for some ξ. Theorem 2.26 follows from this and (2.96).

Chapter 3. DEVELOPMENT OF A GENERAL THEORY OF STATISTICAL DECISION FUNCTIONS

3.1 Formulation of Some Assumptions Regarding the Decision Problem

3.1.1 Assumptions Concerning the Space Ω of Admissible Distribution Functions F

In developing a general theory of decision functions it seems necessary to make some assumptions concerning the space Ω, the weight function $W(F, d^t)$, the space D^t of terminal decisions, the cost function of experimentation, and the decision functions δ at the disposal of the experimenter. The assumptions we shall make are rather weak, and they do not restrict in any serious way the applicability of the theory to problems arising in applications.

In this section we shall formulate some assumptions concerning the space Ω. First we shall introduce some definitions. We shall say that the stochastic process $\{X_i\}$ $(i = 1, 2, \cdots,$ ad inf.) underlying the decision problem is discrete if for any positive integral value r there exists a denumerable subset M_r^* of the r-dimensional sample space M_r such that for all elements F of Ω the probability is 1 that the sample point $(x_1, \cdots, x_r)$ will be an element of M_r^*. Here x_i denotes the observed value of X_i $(i = 1, 2, \cdots, r)$. We shall say that the stochastic process $\{X_i\}$ is absolutely continuous if for any element F of Ω and for any positive integral value r the joint distribution of $X_1, \cdots, X_r$ admits a probability density function.

Assumption 3.1. The stochastic process $\{X_i\}$ $(i = 1, 2, \cdots,$ *ad inf.*) *underlying the decision problem is either discrete or absolutely continuous.*

This assumption is not very restrictive from the point of view of applications, since in most statistical problems arising in practice the stochastic process $\{X_i\}$ will be either discrete or absolutely continuous.

Clearly, if the stochastic process $\{X_i\}$ is discrete, for each positive integral value i there exists a denumerable subset S_i of the real axis such that the probability that the observed value of X_i will fall in S_i is identically equal to 1 for all F. Let the elements of S_i be $a_{i1}, a_{i2}, \cdots$, etc. From the point of view of the theory of statistical decision functions it is immaterial how the various possible values of X_i are labeled. In particular we may put $a_{ij} = j$ $(j = 1, 2, \cdots,$ ad inf.)

without any loss of generality. Thus, in what follows in this chapter we shall assume that in the discrete case the chance variable X_i ($i = 1, 2, \cdots$, ad inf.) can take only positive integral values.

To formulate the next assumption concerning Ω we shall introduce a convergence definition in Ω. For every subset M_r^* of the r-dimensional sample space M_r, let $P(M_r^* \mid F)$ denote the probability, when F is true, that the sample consisting of the observations $x_1, \cdots, x_r$ on $X_1, X_2, \cdots, X_r$, respectively, will be contained in M_r^*. We shall say that F_i converges in the regular sense to F_0 as $i \to \infty$ if for any positive integer r we have

$$\lim_{i=\infty} P(M_r^* \mid F_i) = P(M_r^* \mid F_0) \tag{3.1}$$

uniformly in M_r^*.

The above-defined convergence is called regular in order to distinguish it from other convergence definitions that will be considered later.

Assumption 3.2. *Ω is separable in the sense of the regular convergence definition given in (3.1).*

We shall now show that Assumption 3.2 is a consequence of Assumption 3.1. The reason for formulating both assumptions here is that some results given later in this chapter remain valid when merely Assumption 3.2 is postulated (see, for example, Theorem 3.3).

To prove Assumption 3.2, we shall introduce the following distance definition for each positive integer r: The distance between two elements F_1 and F_2 of Ω is given by

$$t_r(F_1, F_2) = \mathrm{Sup}_{M_r^*} \left| P(M_r^* \mid F_1) - P(M_r^* \mid F_2) \right|$$

where M_r^* may be any subset of the r-dimensional space M_r with the coordinates $x_1, \cdots, x_r$. We shall now show that the separability of Ω in the sense of the convergence definition (3.1) is proved if we can show that Ω is separable in the sense of the metric t_r for each r. For this purpose, suppose that Ω is separable in the sense of the metric t_r for each r. Then for each r there exists a sequence $\{F_{ri}\}$ ($i = 1, 2, \cdots$, ad inf.) of elements of Ω such that $\{F_{ri}\}$ lies dense in Ω in the sense of the metric t_r. Let F be any element of Ω and let $\{\epsilon_r\}$ ($r = 1, 2, \cdots$, ad inf.) be a sequence of positive numbers such that $\lim_{r=\infty} \epsilon_r = 0$. Clearly for each r there exists a positive integer i_r such that $t_r(F, F_{ri_r}) \leqq \epsilon_r$. Since the distance $t_r(F', F'')$ is non-decreasing with increasing r, we see that $\lim_{r=\infty} F_{ri_r} = F$ in the sense of the convergence definition (3.1).

Hence the double sequence $\{F_{ri}\}$ $(r, i = 1, 2, \cdots, \text{ad inf.})$ lies dense in Ω in the sense of the convergence definition (3.1), and the separability of Ω is proved.

We shall now prove the separability of Ω when the underlying stochastic process is discrete. As shown above, it is sufficient to prove the separability of Ω in the sense of the metric t_r for each r. Let Ω_r be the class of all joint distributions of $X_1, \cdots, X_r$ when X_i can take only positive integral values. Clearly the separability of Ω in the sense of the metric t_r is proved, if we show that Ω_r is separable in the sense of the metric t_r. Let Ω_r^* be the set of all elements F of Ω_r which satisfy the following two conditions: (1) for any set of positive integers $c_1, \cdots, c_r$ the probability of the joint event that $X_1 = c_1, X_2 = c_2, \cdots, X_r = c_r$ is a rational number; (2) there exist r integers $u_1, \cdots, u_r$, which may depend on F, such that the probability that $X_i > u_i$ is equal to zero $(i = 1, \cdots, r)$. Clearly Ω_r^* contains only countably many elements and, as can easily be verified, Ω_r^* is a dense subset of Ω_r in the sense of the metric t_r. This completes the proof of Assumption 3.2 when the stochastic process is discrete.

To prove Assumption 3.2 when the underlying stochastic process is absolutely continuous, let Ω_r be the totality of all absolutely continuous distributions of $X_1, \cdots, X_r$. It is sufficient to show that Ω_r is separable in the sense of the metric t_r.

The separability of Ω_r in the sense of the metric t_r can be seen as follows.

Let

$$t_r^*(F_1, F_2) = \int_{M_r} | p(x_1, \cdots, x_r \mid F_1) - p(x_1, \cdots, x_r \mid F_2) | \, dx_1 \cdots dx_r$$

where $p(x_1, \cdots, x_r \mid F)$ denotes the density function in M_r corresponding to the cumulative distribution function F. It is known that Ω_r is separable in the sense of the metric t_r^*.[1] Since $t_r^* \geqq t_r$, the separability of Ω_r in the sense of the metric t_r^* implies the separability of Ω_r in the sense of the metric t_r. This completes the proof of the statement that Assumption 3.2 is a consequence of Assumption 3.1.

3.1.2 Assumptions Concerning the Weight Function $W(F, d^t)$ and the Space D^t of Terminal Decisions

As explained in Section 1.1.5, for any element F of Ω and for any element d^t of D^t the value of $W(F, d^t)$ expresses the loss caused by making the terminal decision d^t when F is the true distribution.

[1] See, for example, Banach [5], pages 12 and 228.

Assumption 3.3. The weight $W(F, d^t)$ is a bounded function of F and d^t.

We shall introduce an intrinsic metric in the space D^t with the help of the weight function $W(F, d^t)$. The (intrinsic) distance between two elements $d_1{}^t$ and $d_2{}^t$ of D^t is defined by the expression

$$R(d_1{}^t, d_2{}^t) = \mathrm{Sup}_F \left| W(F, d_1{}^t) - W(F, d_2{}^t) \right| \tag{3.2}$$

Assumption 3.4. The space D^t is compact in the sense of the metric given in (3.2).

In what follows, by a measurable subset $\bar{D}^t$ of D^t we shall mean an element of the smallest Borel field of subsets of D^t that contains all open subsets of D^t. By a measurable subset of $D = D^t + D^e$ we shall always mean a subset whose intersection with D^t is measurable. Whenever we speak of a subset of D, we shall always mean a measurable subset, even if this is not stated explicitly.

Any finite space D^t is evidently compact. Thus Assumption 3.4 is fulfilled whenever the space D^t is finite. For example, Assumption 3.4 is fulfilled for any problem of testing a hypothesis H, since in this case the space D^t contains only two elements $d_1{}^t$ and $d_2{}^t$, where $d_1{}^t$ denotes the decision to accept H, and $d_2{}^t$ the decision to reject H. There are, however, decision problems treated in the literature which in their conventional form do not fulfill Assumption 3.4. Suppose, for example, that the stochastic process under consideration consists of a single chance variable X_1 and that Ω is the class of all normal distributions with unit variance. Suppose also that the problem is to set up a point estimate for the unknown mean θ of the distribution on the basis of a single observation on X_1. Let $d_{\theta^*}{}^t$ denote the terminal decision to estimate the unknown mean by the value θ^*. Then D^t consists of the elements $d_{\theta^*}{}^t$ corresponding to all possible real values θ^*. Let the weight function $W(\theta, d_{\theta^*}{}^t)$ be equal to $(\theta - \theta^*)^2$. Clearly the space D^t is not compact. It can be made compact, however, by restricting the domain of θ and θ^* to a finite closed interval.

This is not a very serious restriction from the point of view of applications, since in most practical problems we will be able to state a finite closed interval about which we know a priori that it contains the true parameter value θ. The situation will be similar in most of the point estimation and interval estimation problems treated in the literature. If the original conventional form of the problem does not satisfy Assumption 3.4, it will generally be possible to have it fulfilled by restricting the domain of the unknown parameters to a bounded and closed subset of the parameter space.

It would be possible to develop the theory on the basis of a weakened form of Assumption 3.4 which would be fulfilled for estimation problems in their conventional form.[2] However, for the sake of simplicity we shall not attempt to do this. Any weakening of Assumption 3.4 would make the proofs of the main theorems considerably more involved.

3.1.3 Assumptions Concerning the Cost Function of Experimentation

As defined in Section 1.1.5, the symbol

$$c(x; s_1, \cdots, s_k) \tag{3.3}$$

denotes the cost of experimentation when x was the observed sample, the experiment was carried out in k stages, and the ith stage of the experiment consisted of the observations on the chance variables X_j for all j that are elements of s_i. The function $c(x; s_1, \cdots, s_k)$ is defined for any sequence $x = \{x_i\}$ $(i = 1, 2, \cdots,$ ad inf.) for any positive integer k, and for any disjoint and non-empty subsets $s_1, \cdots, s_k$ of the sequence of all positive integers. Of course, the cost $c(x; s_1, \cdots, s_k)$ does not depend on the coordinates x_i of x for which i is not contained in any of the sets $s_1, \cdots, s_k$. If $s_1, \cdots, s_k$ are disjoint subsets of the sequence of positive integers and if s denotes the sequence $\{s_1, \cdots, s_k\}$, then, as stated in Section 1.2.1, the symbol $c(x; s)$ is used as an alternative notation for $c(x; s_1, \cdots, s_k)$; i.e.,

$$c(x; s_1, \cdots, s_k) = c(x; s) \tag{3.4}$$

In what follows, the symbol s will stand for a finite sequence of disjoint and non-empty subsets of the set of all positive integers.

Assumption 3.5. The cost function $c(x; s)$ satisfies the following three conditions:

(i) $c(x; s) \geqq 0$ *for all x and s, and* $c(x; s_1, \cdots, s_k, s_{k+1}) \geqq c(x; s_1, \cdots, s_k)$.

(ii) *For any given s the cost $c(x; s)$ is either a bounded function of x or $c(x; s) = \infty$ identically in x.*

(iii) *There exists a sequence $\{c_m\}$ $(m = 1, 2, \cdots,$ ad inf.) of positive values such that $\lim_{m=\infty} c_m = \infty$ and $c(x; s) \geqq c_m$ for all x, and for all $s = \{s_1, \cdots, s_k\}$ for which the set-theoretical sum of $s_1, \cdots, s_k$ contains at least m elements.*

[2] This was done in a previous publication [70], but the theory developed there is restricted to the special case where the ith stage of the experiment consists of a single observation on X_i.

The reason for admitting the possibility that for some values of s the cost $c(x; s)$ may be equal to ∞ identically in x is that in some situations certain values of s may be practically impossible. For example, if the time needed for carrying out a single stage of the experiment is extremely large, no value $s = \{s_1, \cdots, s_k\}$ with $k > 1$ will be feasible, and this is expressed by putting $c(x; s) = \infty$ for any $k > 1$. It may also happen that an observation on X_j can be made only after the value of X_i has been observed. This can be expressed by putting $c(x; s_1, \cdots, s_k) = \infty$ whenever j is an element of the set-theoretical sum S of $s_1, \cdots, s_k$ and i is not an element of S.

Since the objective of the experimenter is to minimize the risk, assigning the value ∞ to $c(x; s)$ for some values of s, say s^0, is equivalent to restricting the choice of the experimenter to decision functions for which the probability is zero that experimentation will be carried out in accordance with s^0.

The general theory can be reduced to various special cases of interest by imposing some restrictions on the cost function in addition to those listed in Assumption 3.5. For example, if we put $c(x; s_1, \cdots, s_k) = \infty$ for $k > 1$, the general theory reduces to the classical case of experimentation in a single stage. If, in addition, we put $c(x; s_1) = 0$ when s_1 is a subset of the first N integers, and $= \infty$ otherwise, the general theory reduces to the special case where the choice of the experimenter is restricted to decision functions according to which experimentation is carried out in a single stage by observing the values of the first N chance variables $X_1, \cdots, X_N$.

If the cost of experimentation $c(x; s_1, \cdots, s_k)$ depends only on x and the set-theoretical sum of $s_1, \cdots, s_k$, then, for any decision function $\delta(x; s)$ (see Section 1.1.4 for the definition of a decision function), there exists another decision function $\delta^*(x; s)$ such that according to $\delta^*(x; s)$ each stage of the experiment consists of a single observation and

$$r(F, \delta^*) \leqq r(F, \delta)$$

for all F, where $r(F, \delta)$ denotes the risk when F is the true distribution of $X = \{X_i\}$ and δ is the decision function adopted (see Section 1.2.1 for the definition of the risk). Thus, if the cost of experimentation satisfies the above condition, the general theory reduces to the special case where the choice of the experimenter is restricted to decision functions for which each stage of the experiment consists of a single observation. This is the case considered in the recently developed sequential tests of statistical hypotheses (see, for example, [65]).

3.1.4 Assumptions Concerning the Space of Decision Functions at the Disposal of the Experimenter

Before formulating the restrictions to be imposed on the decision functions that can be chosen by the experimenter, we shall introduce a convergence definition in the space of all decision functions. We shall have to treat the discrete case and the absolutely continuous case separately.

If the stochastic process $X = \{X_i\}$ is discrete, we shall say that the sequence $\{\delta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of decision functions converges to the decision function δ_0 as $i \to \infty$, if [3]

$$\lim_{i=\infty} \delta_i(D^* \mid 0) = \delta_0(D^* \mid 0) \tag{3.5}$$

$$\lim_{i=\infty} \delta_i(D^* \mid x; s) = \delta_0(D^* \mid x; s)$$

for any x, s, and for any open subset D^* of D whose boundary has probability zero according to $\delta_0(x; s)$. The topology of the space D is defined as follows: As stated in Section 1.1.4, the space D is the set theoretical sum of D^t and D^e. By the topology of D^t we mean the topology implied by the intrinsic metric introduced in D^t; see equation (3.2). The elements d^e of D^e are to be regarded as discrete points of D. Thus any element d^e is an open subset of D whose boundary is empty.

We could also define convergence in the absolutely continuous case by equation (3.5). This definition of convergence appears, however, to be too strong in the absolutely continuous case, and we shall replace it by a somewhat weaker one. In Section 1.2.1 we introduced the symbol $p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x, \delta)$ [see equation (1.3)] to denote the probability that the experiment is carried out in k stages in accordance with $d_1^e, \cdots, d_k^e$, respectively, and that the terminal decision d^t is an element of the subset $\overline{D}^t$ of D^t when x is the sample point observed and δ is the decision rule adopted. For $k = 0$, the above symbol denotes the probability that no experimentation is made and the terminal decision is an element of $\overline{D}^t$.

Let $p(d_1^e, \cdots, d_k^e \mid x, \delta)$ denote the conditional probability that the experiment is carried out in at least k stages and that the ith stage is carried out in accordance with d_i^e for $i = 1, 2, \cdots, k$ when the sample is known to be equal to x.

For any subset $S = \{i_1, \cdots, i_r\}$ of the set of all positive integers, let R_S denote a subset of the r-dimensional sample space with the

[3] For the definition of the symbols in (3.5), see Sections 1.1.4 and 1.2.1.

coordinates $x_{i_1}, \cdots, x_{i_r}$. We put

$$(3.6)\quad P(d_1^e, \cdots, d_k^e, \bar{D}^t \mid R_S, \delta) = \int_{R_S} p(d_1^e, \cdots, d_k^e, \bar{D}^t \mid x, \delta)\, dx_{i_1} \cdots dx_{i_r}$$

and

$$(3.7)\quad P(d_1^e, \cdots, d_k^e \mid R_S, \delta) = \int_{R_S} p(d_1^e, \cdots, d_k^e \mid x, \delta)\, dx_{i_1} \cdots dx_{i_r}$$

where $S = \{i_1, \cdots, i_r\}$ denotes the set-theoretical sum of $d_1^e, \cdots, d_k^e$ in (3.6), and the set-theoretical sum of $d_1^e, \cdots, d^e{}_{k-1}$ in (3.7). For $k = 0$, the left-hand member of (3.6) is defined to be equal to $\delta(\bar{D}^t \mid 0)$. For $k = 1$, the left-hand member of (3.7) reduces to $\delta(d_1^e \mid 0)$.

If the stochastic process is absolutely continuous, we shall say that

$$(3.8)\quad \lim_{i=\infty} \delta_i = \delta_0$$

if there exists a sequence $\{\bar{D}^t{}_{k_1 \cdots k_m}\}$ $(k_j = 1, \cdots, r_j;\ j = 1, \cdots, m;$ $m = 1, 2, \cdots,$ ad inf.) of subsets of D^t such that

$$(3.9)\quad \lim_{i=\infty} P(d_1^e, \cdots, d_k^e, \bar{D}^t{}_{k_1 \cdots k_m} \mid R_S, \delta_i) = P(d_1^e, \cdots, d_k^e, \bar{D}^t{}_{k_1 \cdots k_m} \mid R_S, \delta_0)$$

and

$$(3.10)\quad \lim_{i=\infty} P(d_1^e, \cdots, d_k^e \mid R_S, \delta_i) = P(d_1^e, \cdots, d_k^e \mid R_S, \delta_0)$$

for any k, $d_1^e, \cdots, d_k^e$, $\bar{D}^t{}_{k_1 \cdots k_m}$, and any bounded set R_S, and the sequence $\{\bar{D}^t{}_{k_1 \cdots k_m}\}$ satisfies the following three conditions:

$$(3.11)\quad \sum_{k=1}^{r_1} \bar{D}_{k_1}{}^t = D^t,\ \sum_{k_m=1}^{r_m} \bar{D}^t{}_{k_1 \cdots k_m} = \bar{D}^t{}_{k_1 \cdots k_{m-1}}$$

$$(3.12)\quad \bar{D}^t{}_{k_1 \cdots k_{m-1}1}, \cdots, \bar{D}^t{}_{k_1 \cdots k_{m-1}r_m} \text{ are disjoint}$$

and

(3.13) Diameter of $\bar{D}^t{}_{k_1 \cdots k_m}$ converges to zero as $m \to \infty$ uniformly in $k_1, \cdots, k_m$

We shall refer to a sequence $\{\bar{D}^t{}_{k_1 \cdots k_m}\}$ of subsets that satisfies (3.11) to (3.13) as a covering net of D^t.

The above-defined convergence in the space of decision functions will be called "regular convergence" (in the discrete as well as the continuous case) to distinguish it from intrinsic convergence, which will be considered later.

Before formulating the restrictions to be imposed on the class of decision functions at the disposal of the experimenter, we want to introduce the notion of convexity. We shall say that a set $\mathfrak{D}$ of decision functions is convex if, for any two elements δ_1 and δ_2 of $\mathfrak{D}$ and for any positive $\alpha < 1$, there exists an element δ of $\mathfrak{D}$ such that the following equations hold for any k, $d_1^e, \cdots, d_k^e$ and any subset $\overline{D}^t$ of D^t:

$$(3.14a)\quad p(d_1^e, \cdots, d_k^e \mid x; \delta) = \alpha p(d_1^e, \cdots, d_k^e \mid x; \delta_1) + (1 - \alpha)p(d_1^e, \cdots, d_k^e \mid x; \delta_2)$$

$$(3.14b)\quad p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x; \delta) = \alpha p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x; \delta_1) + (1 - \alpha)p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x; \delta_2)$$

The above equations imply that δ must satisfy

$$(3.14c)\quad \delta(d^e_{k+1} \mid x; d_1^e, \cdots, d_k^e) = \frac{\alpha p(d_1^e, \cdots, d^e_{k+1} \mid x; \delta_1) + (1 - \alpha)p(d_1^e, \cdots, d^e_{k+1} \mid x; \delta_2)}{\alpha p(d_1^e, \cdots, d_k^e \mid x; \delta_1) + (1 - \alpha)p(d_1^e, \cdots, d_k^e \mid x; \delta_2)}$$

and

$$(3.14d)\quad \delta(\overline{D}^t \mid x; d_1^e, \cdots, d_k^e) = \frac{\alpha p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x; \delta_1) + (1 - \alpha)p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x; \delta_2)}{\alpha p(d_1^e, \cdots, d_k^e \mid x; \delta_1) + (1 - \alpha)p(d_1^e, \cdots, d_k^e \mid x; \delta_2)}$$

provided that $\sum_{i=1}^{2} p(d_1^e, \cdots, d_k^e \mid x; \delta_i) \neq 0$. Thus, for any subset D^* of $D = D^t + D^e$, equations (3.14a) and (3.14b) determine uniquely the value of $\delta(D^* \mid x; d_1^e, \cdots, d_k^e)$, provided that $p(d_1^e, \cdots, d_k^e \mid x; \delta) \neq 0$. If $p(d_1^e, \cdots, d_k^e \mid x; \delta) = 0$, it is irrelevant what value is assigned to $\delta(D^* \mid x; d_1^e, \cdots, d_k^e)$, since it does not influence the risk $r(\xi, \delta)$. Clearly (3.14c) and (3.14d) are not only necessary but also sufficient for the validity (3.14a) and (3.14b).

Obviously the use of a mixed strategy η (probability measure on the space of all admissible decision functions δ) which assigns the probability α to δ_1 and $1 - \alpha$ to δ_2 is equivalent to the use of a pure strategy δ that satisfies (3.14a) and (3.14b). More generally, one can easily verify that any mixed strategy is equivalent to some pure strategy δ.

If the space $\mathfrak{D}$ of decision functions δ at the disposal of the experimenter is convex and closed (in the sense of regular convergence defined before), any discrete mixed strategy η (probability measure η that assigns the probability 1 to some denumerable subset of $\mathfrak{D}$) is equivalent to a pure strategy δ that is an element of $\mathfrak{D}$.

We are now in a position to formulate the assumption we want to make concerning the class $\mathfrak{D}$ of decision functions at the disposal of the experimenter.

Assumption 3.6. The Class $\mathfrak{D}$ of decision functions δ to which the choice of the experimenter is restricted satisfies the following conditions:

(i) *$\mathfrak{D}$ is convex.*

(ii) *$\mathfrak{D}$ is a closed subset of the space of all decision functions in the sense of the regular convergence definition given above.*

(iii) *For any $s = \{s_1, \cdots, s_k\}$ there exists a positive integer c_k depending only on k, such that for any element δ of $\mathfrak{D}$ we have $\delta(d^e \mid x; s) = 0$ for any d^e that is not a subset of the finite set $\{1, 2, \cdots, c_k\}$.*

(iv) *If $c(x; d_1^e, \cdots, d_k^e) = \infty$ identically in x,*

$$p(d_1^e, \cdots, d_k^e \mid x, \delta) = 0$$

for any x and for any element δ of $\mathfrak{D}$.

(v) *A decision function δ is an element of $\mathfrak{D}$ if there exists an element δ_0 of $\mathfrak{D}$ and an element d_0^t of D^t such that for each x and s we have either $\delta(x; s) = \delta_0(x; s)$ or $\delta(d_0^t \mid x; s) = 1$.*

Condition (v) is postulated to insure the possibility of truncation of any element δ of $\mathfrak{D}$. This process of truncation will be used later in the proofs of some lemmas and theorems (see, for example, Lemma 3.2).

The class $\mathfrak{D}_1$ of all decision functions which satisfy conditions (iii) and (iv) of Assumption 3.6 for a given sequence $\{c_k\}$ of integers also satisfies conditions (i), (ii), and (v), as one can easily verify.

The special classes of decision functions mentioned in Section 3.1.3 satisfy Assumption 3.6. In particular, the class of all decision functions according to which experimentation is carried out in one stage by observing the first N chance variables $X_1, \cdots, X_N$ will satisfy Assumption 3.6. This assumption is also fulfilled for the class of all decision functions for which the ith stage of the experiment consists of a single observation on X_i ($i = 1, 2, \cdots$, ad inf.).

We shall now point out the reasons for not imposing the restriction that $\delta(x; s_1, \cdots, s_k) = \delta(x; s'_1, \cdots, s'_r)$ whenever the set-theoretical sum of $s_1, \cdots, s_k$ is equal to that of $s'_1, \cdots, s'_r$. Such a restriction would cause difficulties in the absolutely continuous case due to the assumption that $\mathfrak{D}$ must be a closed subset of the set of all decision functions. We shall define below a sequence $\{\delta_i\}$ ($i = 0, 1, 2, \cdots$, ad inf.) of decision functions such that δ_i satisfies the above restriction for $i \geqq 1$ but δ_0 does not, and $\lim_{i=\infty} \delta_i = \delta_0$ (in the sense of regular

convergence). Let D^t consist of the two elements $d_1{}^t$ and $d_2{}^t$, and let d^{ej} denote the decision to make an observation on X_j. The domain of X_j is restricted to the interval [0, 1]. For $i \geqq 1$, the decision function δ_i is determined by the following equations:

$$\delta_i(d^{e1} \mid 0) = \tfrac{1}{2} \qquad \delta_i(d^{e2} \mid 0) = \tfrac{1}{2}$$

$$\delta_i(d^{e2} \mid x; d^{e1}) = 1 \quad \text{if } \frac{k}{i} \leqq x_1 < \frac{k+1}{i} \text{ for some even } k$$

$$= 0 \quad \text{otherwise}$$

$$\delta_i(d^{e3} \mid x; d^{e1}) = 0 \qquad \delta_i(D^t \mid x; d^{e1}) = 1 - \delta_i(d^{e2} \mid x; d^{e1})$$

$$\delta_i(d^{e1} \mid x; d^{e2}) = 1 \quad \text{if } \frac{k}{i} \leqq x_2 < \frac{k+1}{i} \text{ for some even } k$$

$$= 0 \quad \text{otherwise}$$

$$\delta_i(d^{e3} \mid x; d^{e2}) = 0 \qquad \delta_i(D^t \mid x; d^{e2}) = 1 - \delta_i(d^{e1} \mid x; d^{e2})$$

$$\delta_i(d^{e3} \mid x; d^{e1}, d^{e2}) = 1 \quad \text{if } \frac{k}{i} \leqq x_1 < \frac{k+1}{i} \text{ for even } k \text{ and } \frac{k}{i} \leqq x_2 < \frac{k+1}{i} \text{ for some odd } k$$

$$= 0 \quad \text{otherwise}$$

$$\delta_i(D^t \mid x; d^{e1}, d^{e2}) = 1 - \delta_i(d^{e3} \mid x; d^{e1}, d^{e2})$$

$$\delta_i(d^{e3} \mid x; d^{e2}, d^{e1}) = \delta_i(d^{e3} \mid x; d^{e1}, d^{e2})$$

$$\delta_i(D^t \mid d^{e1}, d^{e2}) = \delta_i(D^t \mid d^{e2}, d^{e1})$$

$$\delta_i(D^t \mid x; d^{e1}, d^{e2}, d^{e3}) = \delta_i(D^t \mid x; d^{e2}, d^{e1}, d^{e3}) = 1$$

and, for any s,

$$\delta_i(d_j{}^t \mid x; s) = \tfrac{1}{2}\delta_i(D^t \mid x; s) \ (j = 1, 2)$$

For any $s = \{s_1, \cdots, s_k\}$ for which $\delta(x; s)$ is not yet determined by the above equations and by the condition that $\delta(x; s_1, \cdots, s_k) = \delta(x; s'_1, \cdots, s'_r)$ if the set-theoretical sum of $s_1, \cdots, s_k$ is equal to that of $s'_1, \cdots, s'_r$, we put

$$\delta_i(d_j{}^t \mid x; s) = \tfrac{1}{2} \ (j = 1, 2)$$

It can easily be seen that δ_i converges to δ_0 as $i \to \infty$, where δ_0 is the decision function determined by the equations

$$\delta_0(d^{e1} \mid 0) = \delta_0(d^{e2} \mid 0) = \tfrac{1}{2} \qquad \delta_0(d^{e2} \mid x; d^{e1}) = \tfrac{1}{2}$$

$$\delta_0(d_j^{\,t} \mid x; d^{e1}) = \tfrac{1}{4}\ (j = 1, 2) \qquad \delta_0(d^{e1} \mid x; d^{e2}) = \tfrac{1}{2}$$

$$\delta_0(d_j^{\,t} \mid x; d^{e2}) = \tfrac{1}{4}\ (j = 1, 2) \qquad \delta_0(d^{e3} \mid x; d^{e1}, d^{e2}) = \tfrac{1}{2}$$

$$\delta_0(d_j^{\,t} \mid x; d^{e1}, d^{e2}) = \tfrac{1}{4}\ (j = 1, 2) \qquad \delta_0(d^{e3} \mid x; d^{e2}, d^{e1}) = 0$$

$$\delta_0(d_j^{\,t} \mid x; d^{e2}, d^{e1}) = \tfrac{1}{2}\ (j = 1, 2) \qquad \delta_0(d_j^{\,t} \mid x; d^{e1}, d^{e2}, d^{e3}) = \tfrac{1}{2}\ (j = 1, 2)$$

The extension of the definition of $\delta_0(x; s)$ for s for which the above equations do not yet determine the value of $\delta_0(x; s)$ can be done in the same way as was done for $\delta_i(x; s)$ $(i \geqq 1)$. Whereas δ_i (for $i \geqq 1$) satisfies the restriction that $\delta_i(x; s_1, \cdots, s_k) = \delta_i(x; s'_1, \cdots, s'_r)$ if the set-theoretical sum of $s_1, \cdots, s_k$ is equal to that of $s'_1, \cdots, s'_r$, the decision function δ_0 does not satisfy this restriction.

3.1.5 Measurability Assumptions

In this section we shall formulate some measurability conditions which will insure the existence of the various integrals that appear in the formulas defining the risk function (see Section 1.2.1).

Let M be the infinite dimensional sample space; i.e., M is the totality of all sequences $x = \{x_i\}$. Let B be the smallest Borel field which contains all sets of points x which are satisfied by the relations

$$x_i < a_i \ (i = 1, 2, \cdots, \text{ad inf.})$$

where the a_i are real numbers or $+\infty$. Furthermore let H be the smallest Borel field of subsets of Ω which contains any subset of Ω that is open in the sense of the convergence definition (3.1). Finally, let T be the smallest Borel field of subsets of $D = D^t + D^e$ which are open in the sense of the topology of the space D defined in Section 3.1.4.

By the symbolic product $H \times T$ we shall mean the smallest Borel field of subsets of the Cartesian product $\Omega \times D$ which contains the Cartesian product of any member of H by any member of T. The symbolic product $H \times B$ is similarly defined.

Only subsets of Ω, D, and M will be considered which are measurable (H), (T), and (B), respectively, even if this is not stated explicitly. The following measurability assumptions are made:

Assumption (μ_1). *$W(F, d^t)$ is a function measurable* $(H \times T)$.

Assumption (μ_2). *For any positive integer* m, $f_m(x_1, \cdots, x_m \mid F)$, *as a function of* x *and* F, *is measurable* $(H \times B)$, *where* $f_m(x_1, \cdots, x_m \mid F)$ *denotes the joint elementary probability law of* $X_1, \cdots, X_m$, *when* F *is the true distribution of* X.[4]

Assumption (μ_3). *For any element* D^* *of the Borel field* T *and for any* $s = \{s_1, \cdots, s_k\}$, *the function* $\delta(D^* \mid x; s)$ *is measurable* (B).

Assumption (μ_4). *For any* $s = \{s_1, \cdots, s_k\}$, *the cost function* $c(x; s)$ *is measurable* (B).

The validity of the above measurability assumptions will be postulated throughout this and subsequent chapters, even if this is not stated explicitly.

We shall now show that the above conditions guarantee that for any δ the risk $r(F, \delta)$ (defined in Section 1.2.1) is a function of F measurable (H). We shall first consider the absolutely continuous case. It follows from Assumption (μ_3) that $p(d_1^e, \cdots, d_k^e, \bar{D}^t \mid x, \delta)$ [defined in (1.3)] is, as a function of x, measurable (B). Hence, because of Assumption (μ_2), $q(d_1^e, \cdots, d_k^e, \bar{D}^t \mid F, \delta)$ [defined in (1.4)] exists and, as a function of F, is measurable (H).[5] From this it follows that $P(\bar{D}^t \mid F, \delta)$ [defined in (1.5)] exists and, as a function of F, is measurable (H).

Because of the compactness of D^t, the integral [see formula (1.6)]

$$r_1(F, \delta) = \int_{D^t} W(F, d^t)\, dP(\bar{D}^t \mid F, \delta)$$

can be represented as the limit of functions $r_{i1}(F, \delta)$ as $i \to \infty$, where

$$r_{i1}(F, \delta) = \sum_{j=1}^{u_i} W(F, d_{ij}{}^t) P(\bar{D}_{ij}{}^t \mid F, \delta)$$

u_i is a finite positive integer, $d_{ij}{}^t$ is an element in $D_{ij}{}^t$, and $\bar{D}_{ij}{}^t$ is an element of T. Since $W(F, d_{ij}{}^t)$ is measurable (H), $r_{i1}(F, \delta)$ is also measurable (H), and therefore $r_1(F, \delta)$ is also measurable (H).

It follows from Assumptions (μ_3) and (μ_4) that the integrand $c(x; d_1^e, \cdots, d_k^e) p(d_1^e, \cdots, d_k^e, D^t \mid x, \delta)$ [see formula (1.7)] as a func-

[4] If the stochastic process is discrete, $f_m(x_1, \cdots, x_m \mid F)$ denotes the probability that $X_1 = x_1$, $X_2 = x_2$, $\cdots$, and $X_m = x_m$. If the stochastic process is absolutely continuous, $f_m(x_1, \cdots, x_m \mid F)$ denotes the density at the point $x_1, \cdots, x_m$.

[5] The integral in (1.4) can be written in the form $\int_M \psi(x, F)\, dB(x)$, and Theorems 9.3 and 9.10 in Saks [44, Chapter III] are applicable. $B(x)$ stands for Borel measure.

tion of x is measurable (B). From this and Assumption (μ_2) it follows that [6]

$$\int_M c(x; d_1^e, \cdots, d_k^e) p(d_1^e, \cdots, d_k^e, D^t \mid x, \delta)\, dF(x)$$

is, as a function of F, measurable (H). Hence $r_2(F, \delta)$ [defined in (1.7)] must be measurable (H). Since $r(F, \delta) = r_1(F, \delta) + r_2(F, \delta)$, the function $r(F, \delta)$ is measurable (H).

The proof that $r(F, \delta)$ is measurable (H) in the discrete case is very similar, except that the integrals in question have to be replaced by sums.

3.2 Weak Intrinsic Compactness of the Space of Decision Functions

3.2.1 Compactness of the Space of Decision Functions in the Sense of Regular Convergence

The main objective of this section is to prove that the space $\mathfrak{D}$ of decision functions is compact in the sense of regular convergence defined in Section 3.1.4. This result will be used in the following section to prove weak intrinsic compactness of the space $\mathfrak{D}$. More precisely, we shall prove the following theorem.

Theorem 3.1. If Assumptions 3.1 to 3.6 hold, the space $\mathfrak{D}$ of decision functions at the disposal of the experimenter is compact in the sense of regular convergence defined in Section 3.1.4.[7]

Proof: First we shall consider the case when the stochastic process $X = \{X_j\}$ $(j = 1, 2, \cdots,$ ad inf.) is discrete. As pointed out in Section 3.1.1, in this case we may assume without loss of generality that for each j the chance variable X_j can take only positive integral values. For any $s = \{s_1, \cdots, s_k\}$, let D_s denote the subset of D consisting of all elements d^t of D^t and all elements d^e of D^e for which d^e is a subset of the set $\{1, 2, \cdots, c_k\}$, where c_s is a positive integer chosen so that condition (iii) of Assumption 3.6 is fulfilled. Clearly $\delta(D_s \mid x; s) = 1$ for any element δ of $\mathfrak{D}$ and for any x and s. Since D_s contains only a finite number of elements outside D^t, it follows from Assumption 3.4 that D_s is compact. Hence, if $\{\delta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) is a sequence of elements of $\mathfrak{D}$, for any given x and s there exist a subsequence $\{i_j\}$ $(j = 1, 2, \cdots,$ ad inf.) of the sequence $\{i\}$ and a probability set

[6] Also this integral can be written in the form $\int_M \psi(x, F)\, dB(x)$.

[7] This theorem is closely related to known theorems on the "weak" compactness of a set of functions. See, for example, Theorem 17b (page 33) of [72].

function $\delta_0(D^* \mid x; s)$ defined for all measurable subsets D^* of D such that

$$\lim_{j=\infty} \delta_{i_j}(D^* \mid x; s) = \delta_0(D^* \mid x; s) \tag{3.15}$$

for any open subset D^* of D whose boundary has probability zero according to the probability set function δ_0. The subsequence $\{i_j\}$ may depend on x and s. However, since there are only denumerably many elements s, and since for any given s the coordinates $x_{i_1}, \cdots, x_{i_r}$ of x on which the value of $\delta(x; s)$ depends can take only denumerably many values, the well-known diagonal procedure can be used to obtain a fixed subsequence $\{i_j\}$ (independent of x and s) for which (3.15) is fulfilled. Thus our theorem is proved in the discrete case.

To prove Theorem 3.1 in the absolutely continuous case, let $\{i_j\}$ ($j = 1, 2, \cdots$, ad inf.) be a subsequence of the sequence $\{i\}$ of positive integers chosen so that, for any $d_1^e, \cdots, d_k^e$ and for any cube T_S with rational vertices in the finite dimensional sample space corresponding to S (S is the set-theoretical sum of $d_1^e, \cdots, d_k^e$), the completely additive set function $P(d_1^e, \cdots, d_k^e, \Delta^t \mid T_S, \delta_{i_j})$ defined for all measurable subsets Δ^t of D^t converges to a completely additive set function $\mu(\Delta^t \mid d_1^e, \cdots, d_k^e, T_S)$; i.e.,

$$\lim_{j=\infty} P(d_1^e, \cdots, d_k^e, Z^t \mid T_S, \delta_{i_j}) = \mu(Z^t \mid d_1^e, \cdots, d_k^e, T_S) \tag{3.16}$$

for any open subset Z^t of D^t for which $\mu(\bar{Z}^t - Z^t \mid d_1^e, \cdots, d_k^e, T_S) = 0$. Since $d_1^e, \cdots, d_k^e$ and T_S can take only denumerably many values, and since the space of all probability measures on a compact metric space is compact (Theorem 2.15), a subsequence $\{i_j\}$ with the above property can be constructed with the help of the diagonal procedure.

For any given subset Δ^t of D^t and for given $d_1^e, \cdots, d_k^e$, the function $\mu(\Delta^t \mid d_1^e, \cdots, d_k^e, T_S)$ defined for all rational cubes T_S can be extended to a completely additive set function $\mu(\Delta^t \mid d_1^e, \cdots, d_k^e, R_S)$ defined for all Borel measurable subsets R_S of the sample space corresponding to S. We shall use the symbol $P(R_S \mid d_1^e, \cdots, d_k^e, \Delta^t)$ synonymously with $\mu(\Delta^t \mid d_1^e, \cdots, d_k^e, R_S)$; i.e., we put

$$P(R_S \mid d_1^e, \cdots, d_k^e, \Delta^t) = \mu(\Delta^t \mid d_1^e, \cdots, d_k^e, R_S) \tag{3.17}$$

This notation will be particularly convenient when we want to keep $\Delta^t, d_1^e, \cdots, d_k^e$ fixed.

We choose the covering net $\{D^t{}_{k_1 \dots k_m}\}$ subject to the following two restrictions:

(*a*) For any element $D^t{}_{k_1 \dots k_m}$ of the covering net we have $D^t{}_{k_1 \dots k_m} \subset \bar{Z}^t{}_{k_1 \dots k_m}$, where $Z_{k_1 \dots k_m}$ denotes the open kernel [8] of $D^t{}_{k_1 \dots k_m}$ and $\bar{Z}^t{}_{k_1 \dots k_m}$ is the closure of $Z_{k_1 \dots k_m}$.

(*b*) $\mu(\bar{Z}^t{}_{k_1 \dots k_m} - Z^t{}_{k_1 \dots k_m} \mid d_1{}^e, \cdots, d_k{}^e, T_S) = 0$ and $P(d_1{}^e, \cdots, d_k{}^e, \bar{Z}^t{}_{k_1 \dots k_m} - Z^t{}_{k_1 \dots k_m} \mid T_S, \delta_{i_j}) = 0$ for any element $D^t{}_{k_1 \dots k_m}$ of the net, for any j, $d_1{}^e, \cdots, d_k{}^e$, and any T_S.

It follows from (3.16), (3.17), and the restriction (*b*) that

$$\lim_{j=\infty} P(d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m} \mid T_S, \delta_{i_j}) = P(T_S \mid d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m}) \tag{3.18}$$

Since the above equation holds for every T_S, it must hold for every bounded and measurable subset R_S; i.e.,

$$\lim_{j=\infty} P(d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m} \mid R_S, \delta_{i_j}) = P(R_S \mid d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m}) \tag{3.19}$$

One can easily verify that the subsequence $\{i_j\}$ can be chosen so that in addition to (3.16) the following relation holds for some set function $P(R_S \mid d_1{}^e, \cdots, d_k{}^e)$:

$$\lim_{j=\infty} P(d_1{}^e, \cdots, d_k{}^e \mid R_S, \delta_{i_j}) = P(R_S \mid d_1{}^e, \cdots, d_k{}^e) \tag{3.20}$$

In this equation, S denotes the set-theoretical sum of $d_1{}^e, \cdots, d^e{}_{k-1}$, while in (3.19) it denotes the sum of $d_1{}^e, \cdots, d_k{}^e$.

We shall now establish some properties of the set functions $P(R_S \mid d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m})$ and $P(R_S \mid d_1{}^e, \cdots, d_k{}^e)$. Clearly

$$P(R_S \mid d_1{}^e, \cdots, d_k{}^e) \geqq 0 \tag{3.21a}$$

$$P(R_S \mid d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m}) \geqq 0$$

$$\sum_{k_m} P(R_S \mid d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_m}) = P(R_S \mid d_1{}^e, \cdots, d_k{}^e, D^t{}_{k_1 \dots k_{m-1}}) \tag{3.21b}$$

Since

$$p(d_1{}^e, \cdots, d^e{}_{k-1}, D^t \mid x, \delta_i) + \sum_{d_k{}^e} p(d_1{}^e, \cdots, d_k{}^e \mid x, \delta_i) = p(d_1{}^e, \cdots, d^e{}_{k-1} \mid x, \delta_i)$$

the above relation remains valid when x is replaced by R_S and p by P. It then follows from (3.19), (3.20), and from the fact that $d_k{}^e$ can take

[8] A point d^t belongs to the open kernel of a subset Δ^t of D^t if and only if there exists a sphere with center d^t and positive radius contained in Δ^t.

only a finite number of values when $d_1^e, \cdots, d^e_{k-1}$ are given [condition (iii) of Assumption 3.6], that

$$(3.21c)\quad P(R_S \mid d_1^e, \cdots, d^e_{k-1}, D^t) + \sum_{d_k^e} P(R_S \mid d_1^e, \cdots, d_k^e) = P(R_S \mid d_1^e, \cdots, d^e_{k-1})$$

Let Z be any element of the sequence $\{Z_{k_1 \cdots k_m}\}$ and let $\{Z^j\}$ be a subsequence of the sequence $\{Z_{k_1 \cdots k_m}\}$ such that the elements of $\{Z^j\}$ are disjoint and $\sum_j \bar{Z}^j = Z$. Then, because of the complete additivity of the set function $\mu(\Delta^t \mid d_1^e, \cdots, d_k^e, T_S)$, we have

$$(3.21d)\quad P(T_S \mid d_1^e, \cdots, d_k^e, Z) = \sum_j P(T_S \mid d_1^e, \cdots, d_k^e, Z^j)$$

For any $d_1^e, \cdots, d_k^e$ and any subset Δ^t of D^t the set functions $P(R_S \mid d_1^e, \cdots, d_k^e, \Delta^t)$ and $P(R_S \mid d_1^e, \cdots, d_k^e)$ are absolutely continuous. Hence, for any $d_1^e, \cdots, d_k^e$ and Δ^t there exists a pair of functions $p^*(x \mid d_1^e, \cdots, d_k^e, \Delta^t)$ and $p(x \mid d_1^e, \cdots, d_k^e)$ such that

$$(3.22)\quad \int_{R_S} p^*(x \mid d_1^e, \cdots, d_k^e, \Delta^t) = P(R_S \mid d_1^e, \cdots, d_k^e, \Delta^t)$$

and

$$(3.23)\quad \int_{R_S} p(x \mid d_1^e, \cdots, d_k^e) = P(R_S \mid d_1^e, \cdots, d_k^e)$$

It follows from (3.21a) to (3.21d) that for almost all x the following conditions hold:

$$(3.24a)\quad p^*(x \mid d_1^e, \cdots, d_k^e, D^t_{k_1 \cdots k_m}) \geqq 0, \qquad p(x \mid d_1^e, \cdots, d_k^e) \geqq 0$$

$$(3.24b)\quad \sum_{k_m} p^*(x \mid d_1^e, \cdots, d_k^e, D^t_{k_1 \cdots k_m}) = p^*(x \mid d_1^e, \cdots, d_k^e, D^t_{k_1 \cdots k_{m-1}})$$

$$(3.24c)\quad p^*(x \mid d_1^e, \cdots, d^e_{k-1}, D^t) + \sum_{d_k^e} p(x \mid d_1^e, \cdots, d_k^e) = p(x \mid d_1^e, \cdots, d^e_{k-1})$$

and

$$(3.24d)\quad p^*(x \mid d_1^e, \cdots, d_k^e, Z) = \sum_j p^*(x \mid d_1^e, \cdots, d_k^e, Z^j)$$

where Z and Z^j are subject to the same conditions as in (3.21d). It follows immediately from restriction (b) that also the condition

$$(3.24e)\quad p^*(x \mid d_1^e, \cdots, d_k^e, D^t_{k_1 \cdots k_m}) = p^*(x \mid d_1^e, \cdots, d_k^e, Z^t_{k_1 \cdots k_m}) = p^*(x \mid d_1^e, \cdots, d_k^e, \bar{Z}^t_{k_1 \cdots k_m})$$

is satisfied for almost all x. One can easily choose the functions p and p^* such that (3.24a) to (3.24e) are satisfied for all x.

For any x, $d_1^e, \cdots, d_k^e$ and for any open subset Z^t of D^t (not necessarily an element of $\{Z^t_{k_1 \cdots k_m}\}$), let

$$(3.25)\quad p(x \mid d_1^e, \cdots, d_k^e, Z^t) = \underset{Z^{*t}}{\text{l.u.b.}}\ p^*(x \mid d_1^e, \cdots, d_k^e, Z^{*t})$$

where Z^{*t} may be the sum of any finite number of elements of the sequence $\{Z^t_{k_1 \cdots k_m}\}$ such that the closure $\bar{Z}^{*t}$ of Z^{*t} is a subset of Z^t. For any x, $d_1^e, \cdots, d_k^e$, the function $p(x \mid d_1^e, \cdots, d_k^e, Z^t)$ can be extended to a completely additive set function $p(x \mid d_1^e, \cdots, d_k^e, \Delta^t)$ defined for all measurable subsets Δ^t of D^t.[9] It follows from (3.24d) and (3.25) that

$$(3.26)\quad p(x \mid d_1^e, \cdots, d_k^e, Z^t_{k_1 \cdots k_m}) = p^*(x \mid d_1^e, \cdots, d_k^e, Z^t_{k_1 \cdots k_m})$$

Since $\sum_{k_1} \cdots \sum_{k_m} p^*(x \mid d_1^e, \cdots, d_k^e, Z^t_{k_1 \cdots k_m}) = p^*(x \mid d_1^e, \cdots, d_k^e, D^t)$ $= p(x \mid d_1^e, \cdots, d_k^e, D^t)$, it follows from (3.26) that $p(x \mid d_1^e, \cdots, d_k^e, \bar{Z}^t_{k_1 \cdots k_m} - Z_{k_1 \cdots k_m}) = 0$. Hence

$$(3.27)\quad p(x \mid d_1^e, \cdots, d_k^e, D^t_{k_1 \cdots k_m}) = p^*(x \mid d_1^e, \cdots, d_k^e, D^t_{k_1 \cdots k_m})$$

Let δ_0 be the function defined by the equations

$$(3.28)\quad \delta_0(\Delta^t \mid x; d_1^e, \cdots, d_k^e) = \frac{p(x \mid d_1^e, \cdots, d_k^e, \Delta^t)}{p(x \mid d_1^e, \cdots, d_k^e)}$$

and

$$(3.29)\quad \delta_0(d^e_{k+1} \mid x; d_1^e, \cdots, d_k^e) = \frac{p(x \mid d_1^e, \cdots, d^e_{k+1})}{p(x \mid d_1^e, \cdots, d_k^e)}$$

If $p(x \mid d_1^e, \cdots, d_k^e) = 0$, we put $\delta_0(d^e_{k+1} \mid x; d_1^e, \cdots, d_k^e) = 0$ and $\delta_0(d^t \mid x; d_1^e, \cdots, d_k^e) = 1$ for some given element d^t of D^t.

It follows from (3.28) that, for any x, $d_1^e, \cdots, d_k^e$, the set function $\delta_0(\Delta^t \mid x; d_1^e, \cdots, d_k^e)$ is non-negative and completely additive. From (3.24c), (3.28), and (3.29) we obtain

$$(3.30)\quad \delta_0(D^t \mid x; d_1^e, \cdots, d_k^e) + \sum_{d^e_{k+1}} \delta_0(d^e_{k+1} \mid x; d_1^e, \cdots, d_k^e) = 1$$

Hence δ_0 is a decision function. Clearly

$$(3.31)\quad p(d_1^e, \cdots, d_k^e \mid x, \delta_0) = p(x \mid d_1^e, \cdots, d_k^e)$$

and

$$(3.31a)\quad p(d_1^e, \cdots, d_k^e, \Delta^t \mid x, \delta_0) = p(x \mid d_1^e, \cdots, d_k^e, \Delta^t)$$

[9] A proof of this is implicit in the proof of Theorem 2.15.

The convergence of δ_{i_j} to δ_0 as $j \to \infty$ is an immediate consequence of the above two equations and equations (3.19), (3.20), (3.22), (3.23), and (3.27). This completes the proof of Theorem 3.1.

3.2.2 Proof of Weak Intrinsic Compactness of the Space of Decision Functions

Let $\{\delta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) be a sequence of decision functions. We shall say that δ_i converges in the intrinsic sense to δ_0 as $i \to \infty$ if

$$\lim_{i=\infty} r(F, \delta_i) = r(F, \delta_0) \tag{3.32}$$

uniformly in F. The above equation implies that

$$\lim_{i=\infty} r(\xi, \delta_i) = r(\xi, \delta_0)$$

uniformly in all a priori distributions ξ. If merely the relation

$$\liminf_{i=\infty} r(\xi, \delta_i) \geqq r(\xi, \delta_0) \tag{3.33}$$

is fulfilled for all ξ, we shall say that δ_i converges weakly to δ_0 in the intrinsic sense.

The space $\mathfrak{D}$ of decision functions at the disposal of the experimenter is said to be compact in the sense of weak intrinsic convergence if, for any sequence $\{\delta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) of elements of $\mathfrak{D}$, there exist an element δ_0 of $\mathfrak{D}$ and a subsequence $\{\delta_{i_j}\}$ $(j = 1, 2, \cdots,$ ad inf.) of the sequence $\{\delta_i\}$ such that

$$\liminf_{j=\infty} r(\xi, \delta_{i_j}) \geqq r(\xi, \delta_0) \tag{3.34}$$

for all ξ. In this section we shall prove the following theorem.

Theorem 3.2. If Assumptions 3.1 to 3.6 hold, regular convergence of δ_i to δ_0 as $i \to \infty$ implies weak intrinsic convergence of δ_i to δ_0. Furthermore, if Assumptions 3.1 to 3.6 hold, the space $\mathfrak{D}$ of decision functions is compact in the sense of weak intrinsic convergence.

Proof: The second half of Theorem 3.2 follows immediately from Theorem 3.1 and the first half of Theorem 3.2. Therefore it is sufficient to prove the first half of Theorem 3.2.

Let $\{\delta_i\}$ be a sequence of decision functions such that $\lim_{i=\infty} \delta_i = \delta_0$ in the regular sense. Also let ξ be any a priori probability measure on Ω. If $\liminf_{i=\infty} r(\xi, \delta_i) = \infty$, the first half of Theorem 3.2 is obviously fulfilled. Therefore it is sufficient to consider probability measures ξ for which $\liminf_{i=\infty} r(\xi, \delta_i) < \infty$. But then we may restrict ourselves to a subse-

quence $\{\delta_{i_j}\}$ of the sequence $\{\delta_i\}$ such that $\lim_{j=\infty} r(\xi, \delta_{i_j}) = \liminf_{i=\infty} r(\xi, \delta_i)$. Therefore, for proving the first half of Theorem 3.2, it is sufficient to consider probability measures ξ for which $r(\xi, \delta_i)$ is a bounded function of i. We shall make this restriction throughout the following proof.

The discrete case: Let $\{\delta_i\}$ be a sequence of elements of $\mathfrak{D}$ and let ξ be an a priori probability measure such that $r(\xi, \delta_i)$ is a bounded function of i ($i \geqq 1$) and δ_i converges to δ_0 as $i \to \infty$ in the regular sense. Because of Theorem 3.1, Theorem 3.2 is proved if we show that

$$\liminf_{i=\infty} r(\xi, \delta_i) \geqq r(\xi, \delta_0) \tag{3.35}$$

Let $\overline{D}^t$ be any open subset of D^t whose boundary has probability zero according to the probability measures $\delta_0(x; d_1^e, \cdots, d_k^e)$ for all $d_1^e, \cdots, d_k^e$, and x. It follows from the regular convergence of δ_i to δ_0 that

$$\lim_{i=\infty} p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x, \delta_i) = p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x, \delta_0) \tag{3.36}$$

Let $f_m(x_1, \cdots, x_m \mid F)$ denote the probability that $X_1 = x_1, \cdots, X_m = x_m$ when F is true. Furthermore let

$$f_m(x_1, \cdots, x_m \mid \xi) = \int_\Omega f_m(x_1, \cdots, x_m \mid F)\, d\xi \tag{3.37}$$

Thus $f_m(x_1, \cdots, x_m \mid \xi)$ is the probability that $X_1 = x_1, \cdots, X_m = x_m$ when ξ is the a priori distribution. Let

$$q(d_1^e, \cdots, d_k^e, \overline{D}^t \mid \xi, \delta) = \sum_{x_1, \cdots, x_m} p(d_1^e, \cdots, d_k^e, \overline{D}^t \mid x, \delta) f_m(x_1, \cdots, x_m \mid \xi) \tag{3.38}$$

where m is a positive integer such that d_i^e is a subset of $\{1, \cdots, m\}$ for $i = 1, \cdots, k$. For $k = 0$, the left-hand member of (3.38) is defined to be equal to $\delta(\overline{D}^t \mid 0)$. It follows from (3.36) that

$$\lim_{i=\infty} q(d_1^e, \cdots, d_k^e, \overline{D}^t \mid \xi, \delta_i) = q(d_1^e, \cdots, d_k^e, \overline{D}^t \mid \xi, \delta_0) \tag{3.39}$$

It follows from condition (iii) of Assumption 3.5 and the boundedness of $r(\xi, \delta_i)$ that for any positive value ρ there exists a positive integer k_ρ, depending only on ρ, such that the probability that the number k of stages of experimentation will not exceed k_ρ is $\geqq 1 - \rho$ when ξ is the a priori distribution in Ω and δ_i is used; i.e.,

$$\sum_{k=0}^{k_\rho} \sum_{d_1^e, \cdots, d_k^e} q(d_1^e, \cdots, d_k^e, D^t \mid \xi, \delta_i) \geqq 1 - \rho \ (i \geqq 1) \tag{3.40}$$

Because of condition (iii) of Assumption 3.6, for any j the set of possible values of d_j^e is finite. From this and relations (3.39) and (3.40) it follows that

$$\sum_{k=0}^{k_\rho} \sum_{d_1^e, \cdots, d_k^e} q(d_1^e, \cdots, d_k^e, D^t \mid \xi, \delta_0) \geqq 1 - \rho \tag{3.41}$$

Since ρ can be chosen arbitrarily small, we obtain from (3.40) and (3.41)

$$\sum_{k=0}^{\infty} \sum_{d_1^e, \cdots, d_k^e} q(d_1^e, \cdots, d_k^e, D^t \mid \xi, \delta_i) = 1 \tag{3.42}$$

$$(i = 0, 1, 2, \cdots, \text{ad inf.})$$

Let

$$r_1(\xi, \delta; d_1^e, \cdots, d_k^e) = \int_\Omega \int_{D^t} W(F, d^t)\, dq(d_1^e, \cdots, d_k^e, \tilde{D}^t \mid F, \delta)\, d\xi \tag{3.43}$$

and let

$$r_2(\xi, \delta; d_1^e, \cdots, d_k^e) \tag{3.44}$$

$$= \sum_{x_1, \cdots, x_m} c(x; d_1^e, \cdots, d_k^e) p(d_1^e, \cdots, d_k^e, D^t \mid x, \delta) f_m(x_1, \cdots, x_m \mid \xi)$$

where m is such that d_i^e is a subset of $\{1, \cdots, m\}$ for $i = 1, 2, \cdots, k$. It follows from (3.42), (3.43), and (3.44) that

$$r(\xi, \delta_i) = \sum_{j=1}^{2} \sum_{k=0}^{\infty} \sum_{d_1^e, \cdots, d_k^e} r_j(\xi, \delta_i; d_1^e, \cdots, d_k^e) \tag{3.45}$$

$$(i = 0, 1, \cdots, \text{ad inf.})$$

Since $W(F, d^t)$ is a continuous function of d^t, and since D^t is compact, it follows from (3.39) that

$$\lim_{i=\infty} \int_{D^t} W(F, d^t)\, dq(d_1^e, \cdots, d_k^e, \tilde{D}^t \mid F, \delta_i) = \int_{D^t} W(F, d^t)\, dq(d_1^e, \cdots, d_k^e, \tilde{D}^t \mid F, \delta_0)$$

Since $W(F, d^t)$ is bounded, the above relation and (3.43) imply that

$$\lim_{i=\infty} r_1(\xi, \delta_i; d_1^e, \cdots, d_k^e) = r_1(\xi, \delta_0; d_1^e, \cdots, d_k^e) \tag{3.46}$$

According to condition (ii) of Assumption 3.5, for any given $d_1^e, \cdots, d_k^e$ and k, $c(x, d_1^e, \cdots, d_k^e)$ is either equal to ∞ identically in x or is a bounded function of x. Since $r(\xi, \delta_i)$ is a bounded function of i, it follows that $p(d_1^e, \cdots, d_k^e, D^t \mid x, \delta_i) = 0$ $(i \geqq 1)$ for any x for which $c(x; d_1^e, \cdots, d_k^e) = \infty$, except perhaps for points $x = (x_1, \cdots, x_m)$ for which $f_m(x_1, \cdots, x_m \mid \xi) = 0$. Hence, because of (3.36), $p(d_1^e, \cdots, d_k^e,$

$D^t \mid x, \delta_0) = 0$ also for any x for which $c(x; d_1^e, \cdots, d_k^e) = \infty$. But then it follows from (3.36) and (3.44) that

$$\lim_{i=\infty} r_2(\xi, \delta_i; d_1^e, \cdots, d_k^e) = r_2(\xi, \delta_0; d_1^e, \cdots, d_k^e) \tag{3.47}$$

Equation (3.35) is an immediate consequence of (3.45), (3.46), and (3.47). This completes the proof of Theorem 3.2 in the discrete case.

The absolutely continuous case: In proving the theorem in the absolutely continuous case, we shall make use of the following lemma.

Lemma 3.1. Let $T_i(S)$ $(i = 0, 1, 2, \cdots, ad\ inf.)$ be a non-negative completely additive set function defined for all measurable subsets S of the r-dimensional sample space M_r. It is assumed that

$$T_i(S) \leqq V(S) \tag{3.48}$$

for all S $(i = 0, 1, 2, \cdots, ad\ inf.)$, where $V(S)$ denotes the Lebesgue measure of S. Let $g(x_1, \cdots, x_r)$ be a non-negative function such that

$$\int_{M_r} g(x_1, \cdots, x_r)\, dx_1 \cdots dx_r < \infty \tag{3.49}$$

Then, if

$$\lim_{i=\infty} T_i(S) = T_0(S) \tag{3.50}$$

we have

$$\lim_{i=\infty} \int_{M_r} g(x_1, \cdots, x_r)\, dT_i = \int_{M_r} g(x_1, \cdots, x_r)\, dT_0 \tag{3.51}$$

Proof: Let $M_{r,c}$ be the sphere in M_r with center at the origin and radius c. Clearly

$$\lim_{c=\infty} \int_{M_{r,c}} g(x_1, \cdots, x_r)\, dx_1 \cdots dx_r = \int_{M_r} g(x_1, \cdots, x_r)\, dx_1 \cdots dx_r \tag{3.52}$$

Hence, because of (3.48), we have

$$\lim_{c=\infty} \left[\int_{M_{r,c}} g(x_1, \cdots, x_r)\, dT_i - \int_{M_r} g(x_1, \cdots, x_r)\, dT_i \right] = 0 \tag{3.53}$$

uniformly in i. Hence our lemma is proved if we show that

$$\lim_{i=\infty} \int_{M_{r,c}} g(x_1, \cdots, x_r)\, dT_i = \int_{M_{r,c}} g(x_1, \cdots, x_r)\, dT_0 \tag{3.54}$$

for any finite c. Let $g_A(x_1, \cdots, x_r) = g(x_1, \cdots, x_r)$, when $g(x_1, \cdots, x_r) \leqq A$, and $= 0$ otherwise. Since

$$\lim_{A=\infty} \int_{M_{r,c}} (g - g_A)\, dx_1 \cdots dx_r = 0$$

it follows from (3.48) that

$$\lim_{A=\infty} \int_{M_{r,c}} (g - g_A)\, dT_i = 0 \tag{3.55}$$

uniformly in i. Hence our lemma is proved if we can show that

$$\lim_{i=\infty} \int_{M_{r,c}} g_A\, dT_i = \int_{M_{r,c}} g_A\, dT_0 \tag{3.56}$$

for any $c > 0$ and any $A > 0$. Let S_j be the set of all points in $M_{r,c}$ for which

$$(j-1)\epsilon \leqq g_A < j\epsilon \tag{3.57}$$

where ϵ is a given positive number. We have

$$\sum_j (j-1)\epsilon \int_{S_j} dT_i \leqq \int_{M_{r,c}} g_A\, dT_i \leqq \sum_j j\epsilon \int_{S_j} dT_i \tag{3.58}$$

$$(i = 0, 1, 2, \cdots, \text{ad inf.})$$

Since, for any ϵ, j can take only a finite number of values, and since ϵ can be chosen arbitrarily small, Lemma 3.1 follows easily from (3.50) and (3.58).

First we shall show that it is sufficient to prove Theorem 3.2 for any finite space D^t. For this purpose, assume that Theorem 3.2 is true for any finite terminal decision space, but that there exist a non-finite compact terminal decision space D^t and a sequence $\{\delta_i\}$ $(i = 0, 1, 2, \cdots, \text{ad inf.})$ of decision functions such that $\lim_{i=\infty} \delta_i = \delta_0$ in the regular sense, and

$$\liminf_{i=\infty} r(\xi, \delta_i) = r(\xi, \delta_0) - \rho \tag{3.59}$$

for some ξ $(\rho > 0)$. Since $\lim_{i=\infty} \delta_i = \delta_0$, there exists a covering net, i.e., a sequence $\{\bar{D}^t_{k_1 \cdots k_m}\}$ $(k_j = 1, \cdots, r_j; j = 1, \cdots, m; m = 1, 2, \cdots,$ ad inf.) of subsets of D^t satisfying the relations (3.11) to (3.13) and

such that (3.9) and (3.10) hold; i.e.,

$$\lim_{i=\infty} P(d_1^e, \cdots, d_k^e, \bar{D}^t_{k_1 \cdots k_m} \mid R_S, \delta_i) = P(d_1^e, \cdots, d_k^e, \bar{D}^t_{k_1 \cdots k_m} \mid R_S, \delta_0) \tag{3.60}$$

and

$$\lim_{i=\infty} P(d_1^e, \cdots, d_k^e \mid R_S, \delta_i) = P(d_1^e, \cdots, d_k^e \mid R_S, \delta_0) \tag{3.61}$$

where S in (3.60) denotes the set-theoretical sum of $d_1^e, \cdots, d_k^e$, and in (3.61) the sum of $d_1^e, \cdots, d^e_{k-1}$. Let m_0 be a fixed value of m, and consider the corresponding finite sequence $\{\bar{D}^t_{k_1 \cdots k_{m_0}}\}$ of subsets of D^t. Let h be the number of elements in this finite sequence. We select one point from each element of the finite sequence $\{\bar{D}^t_{k_1 \cdots k_{m_0}}\}$. Let the points selected be $d_1^t, \cdots, d_h^t$, and let $\bar{D}^t$ denote the set consisting of the points $d_1^t, \cdots, d_h^t$. Let $\bar{\delta}_i$ be the decision rule defined as follows:

$$\begin{aligned} \bar{\delta}_i(d^e \mid x; s) &= \delta_i(d^e \mid x; s) \\ \bar{\delta}_i(d_u{}^t \mid x; s) &= \delta_i(\bar{D}^t_{k_1 \cdots k_{m_0}} \mid x; s) \end{aligned} \qquad (i = 0, 1, 2, \cdots, \text{ad inf.}) \tag{3.62}$$

where $d_u{}^t$ is the element in the sequence $\{d_1^t, \cdots, d_h^t\}$ which is contained in $\bar{D}^t_{k_1 \cdots k_{m_0}}$. Clearly, because of (3.60) and (3.61),

$$\lim_{i=\infty} \bar{\delta}_i = \bar{\delta}_0 \tag{3.63}$$

Given any $\epsilon > 0$, for sufficiently large m_0 we obviously have

$$| r(\xi, \delta_i) - r(\xi, \bar{\delta}_i) | \leqq \epsilon \tag{3.64}$$

for $i = 0, 1, 2, \cdots$, ad inf.

Since for finite D^t our theorem is assumed to be true, we have

$$\liminf_{i=\infty} r(\xi, \bar{\delta}_i) \geqq r(\xi, \bar{\delta}_0) \tag{3.65}$$

Choosing $\epsilon \leqq \rho/3$, we obtain a contradiction from (3.59), (3.64), and (3.65). Thus it is sufficient to prove Theorem 3.2 for finite D^t. In the remainder of the proof we shall assume that D^t consists of the points $d_1^t, \cdots, d_h^t$.

Let $S = i_1, \cdots, i_r$ denote the set-theoretical sum of $d_1^e, \cdots, d_k^e$ and let $f(x; S \mid F)$ denote the marginal joint density function of $X_{i_1}, \cdots, X_{i_r}$ corresponding to the element F of Ω. Then, when ξ is the a priori distribution in Ω and δ is adopted, the probability that the experiment will be carried out in k stages in accordance with $d_1^e, \cdots, d_k^e$, respec-

tively, and that the terminal decision will be equal to $d_u{}^t$ is given by

$$(3.66)\quad q(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid \xi, \delta) = \int_{M_S} p(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid x; \delta) f(x; S \mid \xi)\, dx$$

where M_S denotes the r-dimensional Cartesian space with the coordinates $x_{i_1}, \cdots, x_{i_r}$ and $f(x; S \mid \xi) = \int_\Omega f(x; S \mid F)\, d\xi$. Equation (3.66) can also be written as

$$(3.67)\quad q(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid \xi, \delta) = \int_{M_S} f(x; S \mid \xi)\, dP(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid R_S, \delta)$$

where the set function P is defined in (3.6). Since δ_i, as $i \to \infty$, converges to δ_0 in the regular sense, we have

$$(3.68)\quad \lim_{i=\infty} P(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid R_S, \delta_i) = P(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid R_S, \delta_0)$$

It follows from (3.67), (3.68), and Lemma 3.1 that

$$(3.69)\quad \lim_{i=\infty} q(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid \xi, \delta_i) = q(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid \xi, \delta_0)$$

Similarly to the discrete case, it follows from condition (iii) of Assumption 3.5, from the boundedness of $r(\xi, \delta_i)$ $(i \geqq 1)$, and from equation (3.69) that

$$(3.70)\quad \sum_{k=0}^{\infty} \sum_{d_1{}^e, \cdots, d_k{}^e} q(d_1{}^e, \cdots, d_k{}^e, D^t \mid \xi, \delta_i) = 1 \qquad (i = 0, 1, 2, \cdots, \text{ad inf.})$$

Let

$$(3.71)\quad r_1(\xi, \delta; d_1{}^e, \cdots, d_k{}^e) = \sum_{u=1}^{h} \int_\Omega W(F, d_u{}^t) q(d_1{}^e, \cdots, d_k{}^e, d_u{}^t \mid F, \delta)\, d\xi$$

and

$$(3.72)\quad r_2(\xi, \delta; d_1{}^e, \cdots, d_k{}^e) = \int_{M_S} c(x; d_1{}^e, \cdots, d_k{}^e) p(d_1{}^e, \cdots, d_k{}^e, D^t \mid x, \delta) f(x; S \mid \xi)\, dx$$

where $S = \{i_1, \cdots, i_r\}$ is the set-theoretical sum of $d_1{}^e, \cdots, d_k{}^e$ and

M_S denotes the r-dimensional Cartesian space with the coordinates $x_{i_1}, \cdots, x_{i_r}$. It follows from (3.69) and (3.71) that

$$\lim_{i=\infty} r_1(\xi, \delta_i; d_1^e, \cdots, d_k^e) = r_1(\xi, \delta_0; d_1^e, \cdots, d_k^e) \tag{3.73}$$

Equation (3.72) can be written

$$r_2(\xi, \delta; d_1^e, \cdots, d_k^e) = \int_{M_S} c(x; d_1^e, \cdots, d_k^e) f(x; S \mid \xi)\, dP(d_1^e, \cdots, d_k^e, D^t \mid R_S, \delta) \tag{3.74}$$

It follows from (3.70) that

$$r(\xi, \delta_i) = \sum_{j=1}^{2} \sum_{k=0}^{\infty} \sum_{d_1^e, \cdots, d_k^e} r_j(\xi, \delta_i; d_1^e, \cdots, d_k^e) \qquad (i = 0, 1, 2, \cdots, \text{ad inf.}) \tag{3.75}$$

Because of the regular convergence of δ_i to δ_0, we have

$$\lim_{i=\infty} P(d_1^e, \cdots, d_k^e, D^t \mid R_S, \delta_i) = P(d_1^e, \cdots, d_k^e, D^t \mid R_S, \delta_0) \tag{3.76}$$

We shall now show that

$$\lim_{i=\infty} r_2(\xi, \delta_i; d_1^e, \cdots, d_k^e) = r_2(\xi, \delta_0; d_1^e, \cdots, d_k^e) \tag{3.77}$$

According to condition (ii) of Assumption 3.5, for any given d_1^e, $\cdots$, d_k^e and k, $c(x; d_1^e, \cdots, d_k^e)$ is either equal to ∞ identically in x or is a bounded function of x. Since $r(\xi, \delta_i)$ is a bounded function of i ($i \geqq 1$), it follows that

$$q(d_1^e, \cdots, d_k^e, D^t \mid \xi, \delta_i) = 0 \qquad (i \geqq 1) \tag{3.78}$$

for any $d_1^e, \cdots, d_k^e$ for which $c(x; d_1^e, \cdots, d_k^e) = \infty$ identically in x. Because of (3.69) this remains true also for $i = 0$. Thus in equation (3.75) we can restrict summation with respect to $d_1^e, \cdots, d_k^e$ to values $d_1^e, \cdots, d_k^e$ for which $c(x; d_1^e, \cdots, d_k^e)$ is a bounded function of x. But then it follows from (3.74), (3.76), and Lemma 3.1 that

$$\lim_{i=\infty} r_2(\xi, \delta_i; d_1^e, \cdots, d_k^e) = r_2(\xi, \delta_0; d_1^e, \cdots, d_k^e) \tag{3.79}$$

The first half of Theorem 3.2 is an immediate consequence of (3.75), (3.73), and (3.79). The second half follows from the first half and Theorem 3.1.

The proof of Theorem 3.2 given above in the discrete as well as in the absolutely continuous case shows immediately the validity of the following theorem.

Theorem 3.2a. If Assumptions 3.1 to 3.6 hold, and if only decision functions δ are admitted for which the probability is 1 that the number of stages of the experiment does not exceed a given integer k_0, and if for any s the cost $c(x; s)$ is a bounded function of x, then $\lim_{i=\infty} \delta_i = \delta_0$ in the regular sense implies $\lim_{i=\infty} r(\xi, \delta_i) = r(\xi, \delta_0)$ for all ξ (the convergence is not necessarily uniform in ξ).

3.3 Intrinsic Separability of the Space Ω

For any positive integral value m, let $\mathfrak{D}^m$ denote the set of all decision functions δ which are elements of $\mathfrak{D}$ and have the property that the probability is 1 that experimentation will be carried out in at most m stages when δ is adopted. We shall denote an element of $\mathfrak{D}^m$ by δ^m. For any given m, we shall consider the following four distance definitions in the space Ω.

$$\rho_1(F_1, F_2) = \mathrm{Sup}_R \left| P(R \mid F_1) - P(R \mid F_2) \right| \tag{3.80}$$

where R may be any subset of the m^*-dimensional space of all $(x_1, \cdots, x_{m^*})$ and m^* is a function of m only chosen so that for any element δ^m of $\mathfrak{D}^m$ the probability is zero that an X_i will be observed with $i > m^*$. The existence of a finite m^* with the above property follows from condition (iii) of Assumption 3.6. The symbol $P(R \mid F)$ denotes the probability measure of the set R when F is the true distribution of X.

$$\rho_2(F_1, F_2) = \mathrm{Sup}_{\delta^m} \left| r(F_1, \delta^m) - r(F_2, \delta^m) \right| \tag{3.81}$$

$$\rho_3(F_1, F_2) = \mathrm{Sup}_{d^t} \left| W(F_1, d^t) - W(F_2, d^t) \right| \tag{3.82}$$

$$\rho_4(F_1, F_2) = \rho_1(F_1, F_2) + \rho_3(F_1, F_2) \tag{3.83}$$

We shall call $\rho_2(F_1, F_2)$ the intrinsic distance of F_1 and F_2 relative to $\mathfrak{D}^m$. We shall now prove the following theorem.

Theorem 3.3. If Assumptions 3.2 to 3.6 hold, then for any positive integer m the space Ω is separable in the sense of the intrinsic metric $\rho_2(F_1, F_2)$.

Proof: It follows from Assumption 3.2 that Ω is separable in the sense of the metric $\rho_1(F_1, F_2)$.

We shall now show that Ω is also separable in the sense of the metric $\rho_4(F_1, F_2)$. Since D^t is compact by Assumption 3.4, it follows from Theorem 2.1 that Ω is conditionally compact in the sense of the metric $\rho_3(F_1, F_2)$. Hence, for any $\epsilon > 0$, it is possible to subdivide Ω into a finite number of disjoint subsets $\Omega_1, \cdots, \Omega_r$ such that the diameter of

Ω_i $(i = 1, 2, \cdots, r)$ according to the metric ρ_3 does not exceed ϵ. Since Ω is separable in the sense of the metric ρ_1 there exists a denumerable subset ω_i of Ω_i that lies dense in Ω_i according to the metric ρ_1 $(i = 1, 2, \cdots, r)$. Let ω be the set-theoretical sum of $\omega_1, \cdots, \omega_r$. Clearly ω is denumerable and lies 2ϵ-dense in Ω in the sense of the metric ρ_4. Since ϵ can be chosen arbitrarily small, the separability of Ω in the sense of the metric ρ_4 is proved.

Theorem 3.3 is proved if we can show that, if $\lim_{i=\infty} F_i = F_0$ in the sense of the metric ρ_4, then $\lim_{i=\infty} F_i = F_0$ also in the sense of the metric ρ_2. Let $\{F_i\}$ $(i = 1, 2, \cdots,$ ad inf.) be a sequence for which $\lim_{i=\infty} F_i = F_0$ in the sense of the metric ρ_4. Then

$$\lim_{i=\infty} W(F_i, d^t) = W(F_0, d^t) \tag{3.84}$$

uniformly in d^t, and

$$\lim_{i=\infty} P(R \mid F_i) = P(R \mid F_0) \tag{3.85}$$

uniformly for all subsets R of the m^*-dimensional Cartesian space with the coordinates $x_1, \cdots, x_{m^*}$.

For given values $x_1, \cdots, x_{m^*}$, let $H_i(x_1, \cdots, x_{m^*}, \delta^m)$ denote the conditional expected value of $W(F_i, d^t)$ when δ^m is the decision rule adopted $(i = 0, 1, 2, \cdots,$ ad inf.). Also let $L(x_1, \cdots, x_{m^*}, \delta^m)$ denote the conditional expected cost of experimentation when $x_1, \cdots, x_{m^*}$ are the observed values of $X_1, \cdots, X_{m^*}$ and δ^m is the decision function adopted. Clearly

$$r(F_i, \delta^m) = \int_{M_{m^*}} H_i(x_1, \cdots, x_{m^*}, \delta^m)\, dF_i + \int_{M_{m^*}} L(x_1, \cdots, x_{m^*}, \delta^m)\, dF_i \tag{3.86}$$

where M_{m^*} is the m^*-dimensional Cartesian space with the coordinates $x_1, \cdots, x_{m^*}$. It follows from (3.84) that

$$\lim_{i=\infty} H_i(x_1, \cdots, x_{m^*}, \delta^m) = H_0(x_1, \cdots, x_{m^*}, \delta^m) \tag{3.87}$$

uniformly in $x_1, \cdots, x_{m^*}, \delta^m$. Hence

$$\lim_{i=\infty} \int_{M_{m^*}} [H_i(x_1, \cdots, x_{m^*}, \delta^m) - H_0(x_1, \cdots, x_{m^*}, \delta^m)]\, dF_i = 0 \tag{3.88}$$

uniformly in δ^m. Since H_0 and L are uniformly bounded,[10] it follows

[10] The uniform boundedness of L follows from condition (ii) of Assumption 3.5 and condition (iv) of Assumption 3.6.

from (3.85) that

$$\lim_{i=\infty} \int_{M_m^*} H_0 \, dF_i = \int_{M_m^*} H_0 \, dF_0 \tag{3.89}$$

and

$$\lim_{i=\infty} \int_{M_m^*} L \, dF_i = \int_{M_m^*} L \, dF_0 \tag{3.90}$$

uniformly in δ^m.

Hence we obtain from (3.86), (3.88), (3.89), and (3.90)

$$\lim_{i=\infty} r(F_i, \delta^m) = \int_{M_m^*} H_0 \, dF_0 + \int_{M_m^*} L \, dF_0 = r(F_0, \delta^m) \tag{3.91}$$

uniformly in δ^m. This completes the proof of Theorem 3.3.

3.4 Strict Determinateness of the Decision Problem Viewed as a Zero Sum Two-Person Game

In proving the strict determinateness of the statistical decision problem, we shall make use of the following lemma.

Lemma 3.2. If Assumptions 3.1 to 3.6 hold, for any positive ϵ there exists a positive integer m_ϵ depending only on ϵ, such that

$$\mathrm{Inf}_{\delta^m} r(\xi, \delta^m) \leqq \mathrm{Inf}_{\delta} r(\xi, \delta) + \epsilon \tag{3.92}$$

for any $m \geqq m_\epsilon$ and for any a priori probability distribution ξ in Ω.

Proof: Let n denote the total number of observations made during the course of experimentation and let prob. $\{n \geqq m_\epsilon \mid \xi, \delta\}$ denote the probability that $n \geqq m_\epsilon$ when ξ is the a priori distribution in Ω and δ is the decision function adopted. Let W_0 be an upper bound of $W(F, d^t)$, and let m_ϵ be a positive integer such that

$$c(x; s) \geqq \frac{{W_0}^2}{\epsilon} \tag{3.93}$$

for any x and for any $s = \{s_1, \cdots, s_k\}$ for which $S = s_1 \dot{+} \cdots \dot{+} s_k$ contains at least m_ϵ elements.[11] The existence of such a value m_ϵ follows from condition (iii) of Assumption 3.5.

Let δ_1 be any decision function which is a member of $\mathfrak{D}$. There are two cases to be considered: (*a*) prob. $\{n \geqq m_\epsilon \mid \xi, \delta_1\} \geqq \epsilon/W_0$; (*b*) prob. $\{n \geqq m_\epsilon \mid \xi, \delta_1\} < \epsilon/W_0$. It follows from (3.93) that in case (*a*) we have $r(\xi, \delta_1) \geqq W_0$. In this case, let δ_2 be the rule that we decide on some terminal d^t without taking any observations. Clearly

[11] The symbol $\dot{+}$ stands for "set-theoretical sum."

we shall have $r(\xi, \delta_2) \leqq W_0$, and, therefore, $r(\xi, \delta_2) \leqq r(\xi, \delta_1)$. In case (b), let δ_2 be defined as follows: $\delta_2(x; s) = \delta_1(x; s)$ for any x and for any $s = \{s_1, \cdots, s_k\}$ for which $S = s_1 + \cdots + s_k$ contains less than m_ϵ elements. $\delta_2(d_0^t \mid x; s) = 1$ whenever S contains at least m_ϵ elements, where d_0^t is a fixed element of D^t.[12] Obviously the number of stages of experimentation, when δ_2 is adopted, cannot exceed m_ϵ. Since prob. $\{n \geqq m_\epsilon \mid \xi, \delta_1\} < \epsilon/W_0$, we have

$$r(\xi, \delta_2) \leqq r(\xi, \delta_1) + \epsilon \tag{3.94}$$

Thus Lemma 3.2 is proved.

We are now in a position to prove the following theorem.

Theorem 3.4. If Assumptions 3.1 to 3.6 hold, the decision problem, viewed as a zero sum two-person game, is strictly determined; i.e.,

$$\mathrm{Sup}_\xi \mathrm{Inf}_\delta r(\xi, \delta) = \mathrm{Inf}_\delta \mathrm{Sup}_\xi r(\xi, \delta) \tag{3.95}$$

Proof: It was shown in Chapter 2 that a two-person game is strictly determined if the space A of strategies of the first player is separable in the sense of its intrinsic metric and the space B of strategies of player 2 is weakly compact in the sense of its intrinsic metric (Theorem 2.23). When this result and Theorem 2.24 are applied to the statistical decision problem, it follows from the convexity of $\mathfrak{D}$ [condition (i) of Assumption 3.6] [13] and Theorems 3.2 and 3.3 that

$$\mathrm{Sup}_\xi \mathrm{Inf}_{\delta^m} r(\xi, \delta^m) = \mathrm{Inf}_{\delta^m} \mathrm{Sup}_\xi r(\xi, \delta^m) \tag{3.96}$$

It follows from Lemma 3.2 that for any $\epsilon > 0$ there exists a positive integer m_ϵ such that for $m \geqq m_\epsilon$

$$\mathrm{Sup}_\xi \mathrm{Inf}_\delta r(\xi, \delta) \leqq \mathrm{Sup}_\xi \mathrm{Inf}_{\delta^m} r(\xi, \delta^m) \leqq \mathrm{Sup}_\xi \mathrm{Inf}_\delta r(\xi, \delta) + \epsilon \tag{3.97}$$

From (3.96) and (3.97) we obtain

$$\begin{aligned} \mathrm{Sup}_\xi \mathrm{Inf}_\delta r(\xi, \delta) + \epsilon &\geqq \mathrm{Inf}_{\delta^m} \mathrm{Sup}_\xi r(\xi, \delta^m) \\ &\geqq \mathrm{Inf}_\delta \mathrm{Sup}_\xi r(\xi, \delta) \end{aligned} \tag{3.98}$$

Since ϵ can be chosen arbitrarily small, it follows from (3.98) that

$$\mathrm{Sup}_\xi \mathrm{Inf}_\delta r(\xi, \delta) \geqq \mathrm{Inf}_\delta \mathrm{Sup}_\xi r(\xi, \delta) \tag{3.99}$$

Theorem 3.4 is an immediate consequence of (3.99) and of Lemma 2.3 in Chapter 2.

[12] It follows from condition (v) of Assumption 3.6 that δ_2 is a member of $\mathfrak{D}$.

[13] The convexity of $\mathfrak{D}$ together with condition (ii) of Assumption 3.6 insures that any discrete mixed strategy of the experimenter is equivalent to a pure strategy δ that is an element of $\mathfrak{D}$, as pointed out in Section 3.1.4.

3.5 Theorems on Bayes and Minimax Solutions of the Decision Problem

In this section we shall prove various theorems concerning Bayes and minimax solutions.

Theorem 3.5. If Assumptions 3.1 to 3.6 hold, then for any a priori distribution ξ there exists a decision function δ_ξ such that δ_ξ is a Bayes solution relative to ξ; i.e.,

$$r(\xi, \delta_\xi) = \mathrm{Inf}_\delta\, r(\xi, \delta) \tag{3.100}$$

This theorem is an immediate consequence of Theorems 3.1 and 3.2.

We shall say that $\lim_{i=\infty} \xi_i = \xi_0$ in the ordinary sense if

$$\lim_{i=\infty} \xi_i(\omega) = \xi_0(\omega) \tag{3.101}$$

for any subset ω of Ω which is open in the sense of the intrinsic metric $\rho(F_1, F_2) = \mathrm{Sup}_\delta \left| r(F_1, \delta) - r(F_2, \delta) \right|$ and whose boundary has probability measure zero according to ξ_0.

We shall now prove the following theorem.

Theorem 3.6. Let $\lim_{i=\infty} \xi_i = \xi_0$ in the ordinary sense. Then, if Assumptions 3.1 to 3.6 hold, we have

$$\lim_{i=\infty} \mathrm{Inf}_\delta\, r(\xi_i, \delta) = \mathrm{Inf}_\delta\, r(\xi_0, \delta) \tag{3.102}$$

Proof: For any positive integral value m, let $\mathfrak{D}^m$ be the subset of $\mathfrak{D}$ consisting of those elements δ for which the probability is 1 that experimentation is carried out in at most m stages. Let

$$\rho(F_1, F_2, m) = \mathrm{Sup}_{\delta^m} \left| r(F_1, \delta^m) - r(F_2, \delta^m) \right| \tag{3.103}$$

where δ^m is an element of $\mathfrak{D}^m$. Clearly

$$\rho(F_1, F_2, m) \leqq \rho(F_1, F_2) \tag{3.104}$$

Thus any subset ω of Ω that is open in the sense of the metric $\rho(F_1, F_2, m)$ is open also in the sense of the metric $\rho(F_1, F_2)$. It then follows that $\lim_{i=\infty} \xi_i = \xi_0$ in the ordinary sense also when $\mathfrak{D}$ is replaced by $\mathfrak{D}^m$. According to Theorem 3.3, the space Ω is separable in the sense of the metric $\rho(F_1, F_2, m)$. Hence Theorem 2.14 is applicable[14] and

[14] For the application of Theorem 2.14 to our case it is necessary that $r(F, \delta^m)$ be a bounded function of F and δ^m. But this follows from condition (ii) of Assumption 3.5 and conditions (iii) and (iv) of Assumption 3.6.

we obtain

$$\lim_{i=\infty} r(\xi_i, \delta^m) = r(\xi_0, \delta^m) \tag{3.105}$$

uniformly in δ^m. It follows from the above relation that

$$\lim_{i=\infty} \mathrm{Inf}_{\delta^m} r(\xi_i, \delta^m) = \mathrm{Inf}_{\delta^m} r(\xi_0, \delta^m) \tag{3.106}$$

Theorem 3.6 is an immediate consequence of (3.106) and Lemma 3.2.

Theorem 3.7. If Assumptions 3.1 to 3.6 hold, there exists a minimax solution; i.e., there exists a decision function δ_0 such that

$$\mathrm{Sup}_F r(F, \delta_0) \leqq \mathrm{Sup}_F r(F, \delta) \tag{3.107}$$

for any δ.

Proof: Let $\{\delta_i\}$ $(i = 1, 2, \cdots,$ ad inf.) be a sequence of decision functions such that

$$\lim_{i=\infty} \mathrm{Sup}_F r(F, \delta_i) = \mathrm{Inf}_\delta \mathrm{Sup}_F r(F, \delta) \tag{3.108}$$

It follows from Theorems 3.1 and 3.2 that there exist a subsequence $\{i_j\}$ $(j = 1, 2, \cdots,$ ad inf.) of the sequence $\{i\}$ and a decision function δ_0 such that

$$\liminf_{j=\infty} r(F, \delta_{i_j}) \geqq r(F, \delta_0) \tag{3.109}$$

for all F. Because of (3.108), we have

$$\liminf_{j=\infty} r(F, \delta_{i_j}) \leqq \mathrm{Inf}_\delta \mathrm{Sup}_F r(F, \delta) \tag{3.110}$$

Hence

$$r(F, \delta_0) \leqq \mathrm{Inf}_\delta \mathrm{Sup}_F r(F, \delta)$$

for all F, and therefore

$$\mathrm{Sup}_F r(F, \delta_0) \leqq \mathrm{Inf}_\delta \mathrm{Sup}_F r(F, \delta) \tag{3.111}$$

Obviously the equality sign must hold in the above relation, and Theorem 3.7 is proved.

Theorem 3.8. If Assumptions 3.1 to 3.6 hold, any minimax solution is a Bayes solution in the wide sense.

Proof: Let δ_0 be a minimax solution and $\{\xi_i\}$ a sequence of a priori distributions such that

$$\lim_{i=\infty} \mathrm{Inf}_\delta r(\xi_i, \delta) = \mathrm{Sup}_\xi \mathrm{Inf}_\delta r(\xi, \delta) \tag{3.112}$$

Since δ_0 is a minimax solution, we have

$$\text{Sup}_F\, r(F, \delta_0) = \text{Inf}_\delta \text{Sup}_\xi\, r(\xi, \delta) \tag{3.113}$$

Hence, because of Theorem 3.4, we have

$$\text{Sup}_F\, r(F, \delta_0) = \lim_{i=\infty} \text{Inf}_\delta\, r(\xi_i, \delta) \tag{3.114}$$

and therefore

$$r(\xi_j, \delta_0) \leqq \lim_{i=\infty} \text{Inf}_\delta\, r(\xi_i, \delta) \tag{3.115}$$

Theorem 3.8 is an immediate consequence of (3.115).

Theorem 3.9. If Assumptions 3.1 to 3.6 hold, and if ξ_0 is a least favorable a priori distribution, then any minimax solution is also a Bayes solution relative to ξ_0.

Proof: Let ξ_0 be a least favorable a priori distribution; i.e., ξ_0 satisfies the relation

$$\text{Inf}_\delta\, r(\xi_0, \delta) = \text{Sup}_\xi \text{Inf}_\delta\, r(\xi, \delta) \tag{3.116}$$

Let δ_0 be a minimax solution. Then

$$\text{Sup}_F\, r(F, \delta_0) = \text{Inf}_\delta \text{Sup}_\xi\, r(\xi, \delta) \tag{3.117}$$

Hence, because of Theorem 3.4,

$$r(\xi_0, \delta_0) \leqq \text{Sup}_F\, r(F, \delta_0) = \text{Inf}_\delta\, r(\xi_0, \delta) \tag{3.118}$$

and our theorem is proved.

Theorem 3.10. Let ξ_0 be a least favorable a priori distribution, δ_0 a minimax solution, and ω the set of all elements F of Ω for which

$$r(F, \delta_0) < \text{Sup}_F\, r(F, \delta_0)$$

Then, if Assumptions 3.1 to 3.6 hold, $\xi_0(\omega) = 0$.

Proof: According to equation (3.118) we have

$$\text{Sup}_F\, r(F, \delta_0) = \text{Inf}_\delta\, r(\xi_0, \delta) \tag{3.119}$$

Clearly the above equation implies that

$$\text{Sup}_F\, r(F, \delta_0) = r(\xi_0, \delta_0) \tag{3.120}$$

But (3.120) can hold only if $\xi_0(\omega) = 0$, and our theorem is proved.

We shall say that an element F of Ω is degenerate relative to the a priori distribution ξ if there exists a subset ω of Ω such that ω contains

F, ω is open in the sense of the intrinsic metric $\rho(F_1, F_2) = \mathrm{Sup}_\delta \left| r(F_1, \delta) - r(F_2, \delta) \right|$, and $\xi(\omega) = 0$.

Theorem 3.11. If ξ_0 is a least favorable a priori distribution and δ_0 is a minimax solution, and if Assumptions 3.1 to 3.6 hold, then

$$r(F, \delta_0) = \mathrm{Max}_F\, r(F, \delta_0) \tag{3.121}$$

for all F which are not degenerate relative to ξ_0.

Proof: Suppose that there exists an element F_0 such that F_0 is not degenerate relative to ξ_0 and

$$r(F_0, \delta_0) < \mathrm{Sup}_F\, r(F, \delta_0) \tag{3.122}$$

Then there exists an open subset ω [in the sense of the intrinsic metric $\rho(F_1, F_2)$] that contains F_0 and such that

$$r(F, \delta_0) < \mathrm{Sup}_F\, r(F, \delta_0) \tag{3.123}$$

for all F in ω. Since F_0 is not degenerate, we have

$$\xi_0(\omega) > 0 \tag{3.124}$$

Equations (3.123) and (3.124) contradict Theorem 3.10. Hence (3.122) is impossible and our theorem is proved.

We shall now show that there exists a minimax solution which is a limit of a sequence of Bayes solutions in the strict sense. For this purpose, we shall need the lemmas stated and proved below.

We shall say that a decision function δ_1 is obtained from the decision function δ_0 by truncation after the mth stage of the experiment if

$$\delta_1(x; d_1^e, \cdots, d_k^e) = \delta_0(x; d_1^e, \cdots, d_k^e) \tag{3.125}$$

for any $k < m$ and if

$$\delta_1(D^t \mid x; d_1^e, \cdots, d_m^e) = 1 \tag{3.126}$$

Lemma 3.3. If δ_0^m is a decision function obtained from δ_0 by truncation after the mth stage, and if Assumptions 3.1 to 3.6 hold, then

$$\lim_{m=\infty} r(\xi, \delta_0^m) = r(\xi, \delta_0) \tag{3.127}$$

Proof: If the probability is positive that the experimentation will go on indefinitely when δ_0 is adopted and ξ is the a priori probability measure, then $r(\xi, \delta_0) = \infty$ and $\lim_{m=\infty} r(\xi, \delta_0^m) = \infty$. Thus it is sufficient to consider the case when the probability in question is zero.

Let $r_1(\xi, \delta, d_1^e, \cdots, d_k^e)$ and $r_2(\xi, \delta, d_1^e, \cdots, d_k^e)$ be defined as in (3.43) and (3.44).[15] We have

$$r(\xi, \delta_0) = \sum_{j=1}^{2} \sum_{k=0}^{\infty} \sum_{d_1^e, \cdots, d_k^e} r_j(\xi, \delta_0, d_1^e, \cdots, d_k^e) \tag{3.128}$$

Clearly

$$\sum_{j=1}^{2} \sum_{k=0}^{m-1} \sum_{d_1^e, \cdots, d_k^e} r_j(\xi, \delta_0, d_1^e, \cdots, d_k^e) \leqq r(\xi, \delta_0^m) \leqq \sum_{j=1}^{2} \sum_{k=0}^{m} \sum_{d_1^e, \cdots, d_k^e} r_j(\xi, \delta_0, d_1^e, \cdots, d_k^e) + P_m W_0 \tag{3.129}$$

where W_0 is an upper bound of $W(F, d^t)$ and P_m is the probability that experimentation will be carried out in at least m stages when δ_0 is adopted and ξ is the a priori probability measure. Since the probability is zero that experimentation will go on indefinitely, we have

$$\lim_{m=\infty} P_m = 0 \tag{3.130}$$

Lemma 3.3 follows from (3.128), (3.129), and (3.130).

Lemma 3.4. Let $\{\xi_i\}$ $(i = 0, 1, 2, \cdots,$ *ad inf.*$)$ *be a sequence of a priori probability measures such that*

$$\lim_{i=\infty} \operatorname{Sup}_{\delta^m} \left| r(\xi_i, \delta^m) - r(\xi_0, \delta^m) \right| = 0 \tag{3.131}$$

for $m = 1, 2, 3, \cdots,$ *ad inf. Then, if Assumptions 3.1 to 3.6 hold, fo. any decision function* δ_0 *we have*

$$\liminf_{i=\infty} r(\xi_i, \delta_0) \geqq r(\xi_0, \delta_0) \tag{3.132}$$

Proof: Let $\{\xi_i\}$ $(i = 0, 1, 2, \cdots,$ ad inf.$)$ be a sequence of probability measures for which the conditions of Lemma 3.4 are fulfilled. Let δ_0 be a decision function and let δ_0^m be a decision function obtained from δ_0 by truncation after the mth stage of the experiment. It follows from (3.131) that

$$\lim_{i=\infty} r(\xi_i, \delta_0^m) = r(\xi_0, \delta_0^m) \tag{3.133}$$

If $\liminf_{i=\infty} r(\xi_i, \delta_0) = \infty$, Lemma 3.4 obviously holds. Therefore it is sufficient to consider the case when $\liminf_{i=\infty} r(\xi_i, \delta_0) < \infty$. Let $\{i_j\}$

[15] Equations (3.43) and (3.44) refer to the discrete case. It is clear what the corresponding formulas are in the absolutely continuous case.

($j = 1, 2, \cdots$, ad inf.) be a subsequence of the sequence $\{i\}$ such that

$$\lim_{j=\infty} r(\xi_{i_j}, \delta_0) = \liminf_{i=\infty} r(\xi_i, \delta_0) < \infty \tag{3.134}$$

Let P_{jm} be the probability that the experiment will be carried out in at least m stages when δ_0 is adopted and ξ_{i_j} is the a priori probability measure. Since $r(\xi_{i_j}, \delta_0)$ is a bounded function of j, we have

$$\lim_{m=\infty} P_{jm} = 0 \tag{3.135}$$

uniformly in j. Hence, for any $\epsilon > 0$, there exists a positive integer m_ϵ, depending only on ϵ, such that

$$r(\xi_{i_j}, \delta_0{}^m) \leqq r(\xi_{i_j}, \delta_0) + \epsilon \tag{3.136}$$

for all $m \geqq m_\epsilon$. From (3.133) and (3.136) it follows that

$$\lim_{j=\infty} r(\xi_{i_j}, \delta_0) \geqq r(\xi_0, \delta_0{}^m) - \epsilon \tag{3.137}$$

for all $m \geqq m_\epsilon$. Thus, because of Lemma 3.3, we have

$$\lim_{j=\infty} r(\xi_{i_j}, \delta_0) \geqq r(\xi_0, \delta_0) - \epsilon \tag{3.138}$$

Since the above equation holds for any $\epsilon > 0$, Lemma 3.4 is proved.

Lemma 3.5. Let $\{\xi_i\}$ ($i = 0, 1, 2, \cdots$, ad inf.) be a sequence of a priori probability measures such that $\lim_{i=\infty} \xi_i(\omega) = \xi_0(\omega)$ for any open set ω whose boundary has probability zero according to ξ_0. The terms "open" and "boundary" are meant here in the sense of the following convergence definition in Ω: F_i converges to F_0 as $i \to \infty$ if $\lim_{i=\infty} F_i = F_0$ in the sense of regular convergence [see equation (3.1)] and if $\lim_{i=\infty} W(F_i, d^t) = W(F_0, d^t)$ uniformly in d^t. Then, if Assumptions 3.1 to 3.6 hold, the sequence $\{\xi_i\}$ ($i = 0, 1, 2, \cdots$, ad inf.) satisfies the condition (3.131) of the preceding lemma.

Proof: Let $\{\xi_i\}$ ($i = 0, 1, 2, \cdots$, ad inf.) be a sequence of probability measures which satisfies the assumption of Lemma 3.5. For any positive integral value m, let

$$\rho(F_1, F_2, m) = \mathrm{Sup}_{\delta^m} \left| r(F_2, \delta^m) - r(F_1, \delta^m) \right| \tag{3.139}$$

and

$$\rho(\xi', \xi'', m) = \mathrm{Sup}_{\delta^m} \left| r(\xi', \delta^m) - r(\xi'', \delta^m) \right| \tag{3.140}$$

In proving Theorem 3.3, we have shown that convergence in the sense of the definition given in Lemma 3.5 implies convergence in the sense of the metric (3.139) for any m. Let m_0 be a positive integer

and ω_0 be a subset of Ω such that ω_0 is open in the sense of the metric $\rho(F_1, F_2, m_0)$ and the boundary of ω_0 [in the sense of the metric $\rho(F_1, F_2, m_0)$] has probability zero according to ξ_0. Since convergence in the sense of the definition in Lemma 3.5 implies convergence in the sense of the metric $\rho(F_1, F_2, m_0)$, it follows that ω_0 is open in the sense of Lemma 3.5, and the boundary of ω_0 in the sense of Lemma 3.5 is a subset of the boundary of ω_0 in the sense of the metric $\rho(F_1, F_2, m_0)$. Thus, because of the assumption of Lemma 3.5, we have $\lim_{i=\infty} \xi_i(\omega_0) = \xi_0(\omega_0)$. More generally, $\lim_{i=\infty} \xi_i(\omega) = \xi_0(\omega)$ if there exists a positive integer m such that ω is open in the sense of the metric $\rho(F_1, F_2, m)$ and the boundary of ω [in the sense of $\rho(F_1, F_2, m)$] has probability zero according to ξ_0. Lemma 3.5 follows from this and from Theorem 2.14 in Chapter 2.

Lemma 3.6. If Assumptions 3.1 to 3.6 hold, there exists a fixed sequence $\{\xi_i\}$ ($i = 1, 2, \cdots$, *ad inf.*) *of probability measures such that for any positive integer* m *the sequence* $\{\xi_i\}$ *lies dense in the space of all* ξ*'s in the sense of the metric (3.140).*

Proof: According to Theorem 3.3, Ω is separable in the sense of the metric (3.139). It then follows from Theorem 2.16 that the space of all ξ's is separable in the sense of (3.140). Hence, for any positive integer m, there exists a sequence $\{\xi_{im}\}$ ($i = 1, 2, \cdots$, ad inf.) that is dense in the space of all ξ's in the sense of the metric (3.140). A sequence $\{\xi_i\}$ that contains every ξ_{im} as an element obviously satisfies Lemma 3.6.

Theorem 3.12. Let $\{\xi_i\}$ ($i = 1, 2, \cdots$, *ad inf.*) *be a fixed sequence of probability measures on* Ω *such that for any positive integer* m *the sequence* $\{\xi_i\}$ *is dense in the space consisting of all* ξ_i ($i = 1, 2, \cdots$) *and of all* ξ_F (F *may be any element of* Ω) *in the sense of the metric (3.140), where* ξ_F *denotes the probability measure that assigns the probability 1 to* F*. Then, if Assumptions 3.1 to 3.6 hold, there exist a minimax solution* δ_0 *and a sequence* $\{\delta_j\}$ ($j = 1, 2, \cdots$, *ad inf.*) *of decision functions such that* $\lim_{j=\infty} \delta_j = \delta_0$ *and, for each* j, δ_j *is a Bayes solution relative to some probability measure* ξ'_j *that is a linear combination of a finite number of elements of the sequence* $\{\xi_i\}$.

Proof: Let δ_j be a minimax solution of the decision problem when the choice of ξ is restricted to linear combinations of $\xi_1, \cdots, \xi_j$ with non-negative coefficients; i.e., δ_j satisfies the condition

$$\text{(3.141)} \qquad \operatorname*{Max}_{i \leqq j} r(\xi_i, \delta_j) \leqq \operatorname*{Max}_{i \leqq j} r(\xi_i, \delta)$$

for any δ. The restriction imposed on the choice of ξ makes the decision problem equivalent with a two-person game where the possible pure strategies of the first player are represented by $\xi_1, \cdots, \xi_j$. Thus δ_j must be a Bayes solution relative to some linear combination ξ'_j of $\xi_1, \cdots, \xi_j$. Let δ_0 be the limit of a convergent subsequence of $\{\delta_j\}$. It then follows from Theorem 3.2 and (3.141) that [16]

$$\text{Sup}_i\, r(\xi_i, \delta_0) \leqq \text{Sup}_i\, r(\xi_i, \delta) \tag{3.142}$$

for any δ. It follows from Lemma 3.4 that

$$\text{Sup}_i\, r(\xi_i, \delta) = \text{Sup}_F\, r(F, \delta) \tag{3.143}$$

Thus δ_0 is a minimax solution when no restriction is imposed on the choice of ξ, and our theorem is proved.

It may be of interest to mention some particular possibilities for the choice of a sequence $\{\xi_i\}$ that lies dense in the space consisting of all ξ_i and all ξ_F in the sense of the metric in (3.140) for all m. Let $\{F_i\}$ $(i = 1, 2, \cdots,$ ad inf.) be a sequence of elements of Ω such that for any positive integer m the sequence $\{F_i\}$ lies dense in Ω in the sense of the metric (3.139). Let ξ_i be the probability measure that assigns the probability 1 to F_i. Then $\{\xi_i\}$ is dense in the set of all ξ_F's in the sense of the metric (3.140) for all m. It is also possible to choose a sequence $\{\xi_i\}$ $(i = 1, 2, \cdots,$ ad inf.) such that $\xi_i(F_j) > 0$ for all j, $\sum_{j=1}^{\infty} \xi_i(F_j) = 1$, and $\{\xi_i\}$ lies dense in the space consisting of all ξ_i and ξ_F's in the sense of the metric (3.140) for all m.

We shall now formulate an additional assumption which will make it possible to prove some stronger theorems.

Assumption 3.7. *The space Ω is compact in the sense of regular convergence defined in Section 3.1.1. If $\lim_{i=\infty} F_i = F_0$ in the regular sense, then*

$$\lim_{i=\infty} W(F_i, d^t) = W(F_0, d^t) \tag{3.144}$$

uniformly in d^t.

Theorem 3.13. Let $\{\xi_i\}$ $(i = 0, 1, 2, \cdots,$ ad inf.$)$ be a sequence of a priori probability measures such that $\lim_{i=\infty} \xi_i(\omega) = \xi_0(\omega)$ for any open subset ω (in the sense of regular convergence in Ω) whose boundary (in the sense of regular convergence in Ω) has probability zero according to ξ_0. Then, if Assumptions 3.1 to 3.7 hold,

$$\text{(i)} \qquad \lim_{i=\infty} r(\xi_i, \delta^m) = r(\xi_0, \delta^m)$$

[16] The argument in showing (3.142) is essentially the same as that used in proving Theorem 2.23.

uniformly in δ^m for any m.

(ii) $$\lim_{i=\infty} \mathrm{Inf}_\delta\, r(\xi_i, \delta) = \mathrm{Inf}_\delta\, r(\xi_0, \delta)$$

(iii) $$\liminf_{i=\infty} r(\xi_i, \delta) \geqq r(\xi_0, \delta)$$

Proof: Let $\{\xi_i\}$ be a sequence of probability measures satisfying the condition of our theorem. Statement (i) is an immediate consequence of Assumption 3.7 and Lemma 3.5. It implies that

(3.145) $$\lim_{i=\infty} \mathrm{Inf}_{\delta^m}\, r(\xi_i, \delta^m) = \mathrm{Inf}_{\delta^m}\, r(\xi_0, \delta^m)$$

Statement (ii) is an immediate consequence of (3.145) and Lemma 3.2. Statement (iii) follows from Statement (i) and Lemma 3.4.

Theorem 3.14. If Assumptions 3.1 to 3.7 hold, there exists a least favorable a priori distribution.

Proof: Let $\{\xi_i\}$ be a sequence of probability measures such that

(3.146) $$\lim_{i=\infty} \mathrm{Inf}_\delta\, r(\xi_i, \delta) = \mathrm{Sup}_\xi\, \mathrm{Inf}_\delta\, r(\xi, \delta)$$

It follows from Theorem 2.15 that there exists a subsequence $\{i_j\}$ ($j = 1, 2, \cdots$, ad inf.) of the sequence $\{i\}$ and a probability measure ξ_0 such that

(3.147) $$\lim_{j=\infty} \xi_{i_j}(\omega) = \xi_0(\omega)$$

for any open set ω (in the sense of regular convergence in Ω) whose boundary has probability zero according to ξ_0. Hence, because of Theorem 3.13,

(3.148) $$\lim_{j=\infty} \mathrm{Inf}_\delta\, r(\xi_{i_j}, \delta) = \mathrm{Inf}_\delta\, r(\xi_0, \delta)$$

It follows from (3.146) and (3.148) that ξ_0 is a least favorable a priori distribution, and our theorem is proved.

Theorem 3.15. If Assumptions 3.1 to 3.7 hold and if δ_0 is a Bayes solution in the wide sense, δ_0 is also a Bayes solution in the strict sense.

Proof: Let δ_0 be a Bayes solution relative to the sequence $\{\xi_i\}$ ($i = 1, 2, \cdots$, ad inf.) of probability measures. Let $\{\xi'_i\}$ ($i = 1, 2, \cdots$, ad inf.) be a subsequence of the sequence $\{\xi_i\}$ and ξ_0 a probability measure such that $\lim_{i=\infty} \xi'_i(\omega) = \xi_0(\omega)$ for any open set ω (in the sense of regular convergence) whose boundary has probability zero according to ξ_0. It follows from Theorem 3.13 that

(3.149) $$\lim_{i=\infty} \mathrm{Inf}_\delta\, r(\xi'_i, \delta) = \mathrm{Inf}_\delta\, r(\xi_0, \delta)$$

Since δ_0 is a Bayes solution relative to the sequence ξ'_i, we have

$$\lim_{i=\infty} [r(\xi'_i, \delta_0) - \mathrm{Inf}_\delta\, r(\xi'_i, \delta)] = 0 \tag{3.150}$$

From (3.149) and (3.150) we obtain

$$\lim_{i=\infty} r(\xi'_i, \delta_0) = \mathrm{Inf}_\delta\, r(\xi_0, \delta) \tag{3.151}$$

From statement (iii) of Theorem 3.13 it follows that

$$\lim_{i=\infty} r(\xi'_i, \delta_0) \geqq r(\xi_0, \delta_0) \tag{3.152}$$

Theorem 3.15 is an immediate consequence of (3.151) and (3.152).

Theorem 3.16. If Assumptions 3.1 to 3.7 hold, a limit of a sequence of Bayes solutions in the strict sense is itself a Bayes solution in the strict sense.

Proof: Let $\{\xi_i\}$ ($i = 1, 2, \cdots$, ad inf.) be a sequence of probability measures, and $\{\delta_i\}$ ($i = 0, 1, 2, \cdots$, ad inf.) a sequence of decision functions such that, for each $i > 0$, δ_i is a Bayes solution relative to ξ_i and $\lim_{i=\infty} \delta_i = \delta_0$. There exists a subsequence $\{i_j\}$ ($j = 1, 2, \cdots$, ad inf.) of the sequence $\{i\}$ and a probability measure ξ_0 such that $\lim_j \xi_{i_j}(\omega) = \xi_0(\omega)$ for any open set ω (in the sense of regular convergence in Ω) whose boundary has probability zero according to ξ_0. It then follows from Theorem 3.13 that

$$\lim_{j=\infty} r(\xi_{i_j}, \delta_{i_j}) = \mathrm{Inf}_\delta\, r(\xi_0, \delta) \tag{3.153}$$

Let δ_i^m be the decision function determined as follows:

$$\delta_i^m(x; d_1^e, \cdots, d_k^e) = \delta_i(x; d_1^e, \cdots, d_k^e) \tag{3.154}$$

for $k < m$, and

$$\delta_i^m(d_0^t \mid x; d_1^e, \cdots, d_m^e) = 1 \tag{3.155}$$

where d_0^t is a fixed element of D^t. Let P_{jm} denote the probability that experimentation will consist of at least m stages when ξ_{i_j} is the a priori distribution and δ_{i_j} is adopted. Since $r(\xi_{i_j}, \delta_{i_j})$ is a bounded function of j, we have

$$\lim_{m=\infty} P_{jm} = 0 \tag{3.156}$$

uniformly in j. From this it follows that for any $\epsilon > 0$ there exists a positive integer m_ϵ, depending only on ϵ, such that

$$r(\xi_{i_j}, \delta_{i_j}^m) \leqq r(\xi_{i_j}, \delta_{i_j}) + \epsilon \tag{3.157}$$

for all $m \geqq m_\epsilon$. From this and (3.153) we obtain

$$\limsup_{j=\infty} r(\xi_{i_j}, \delta_{i_j}^m) \leqq \mathrm{Inf}_\delta\, r(\xi_0, \delta) + \epsilon \tag{3.158}$$

for $m \geqq m_\epsilon$. According to Theorem 3.13, we have

$$\lim_{j=\infty} [r(\xi_{i_j}, \delta_{i_j}^m) - r(\xi_0, \delta_{i_j}^m)] = 0 \tag{3.159}$$

Hence

$$\limsup_{j=\infty} r(\xi_0, \delta_{i_j}^m) \leqq \mathrm{Inf}_\delta\, r(\xi_0, \delta) + \epsilon \tag{3.160}$$

for $m \geqq m_\epsilon$. Clearly

$$\lim_{j=\infty} \delta_{i_j}^m = \delta_0^m \tag{3.161}$$

Thus, because of Theorem 3.2, we have

$$\liminf_{i=\infty} r(\xi_0, \delta_{i_j}^m) \geqq r(\xi_0, \delta_0^m) \tag{3.162}$$

From (3.160) and (3.162) it follows that

$$r(\xi_0, \delta_0^m) \leqq \mathrm{Inf}_\delta\, r(\xi_0, \delta) + \epsilon \tag{3.163}$$

for $m \geqq m_\epsilon$. According to Lemma 3.3, we have

$$\lim_{m=\infty} r(\xi_0, \delta_0^m) = r(\xi_0, \delta_0) \tag{3.164}$$

Equations (3.163) and (3.164) imply that $r(\xi_0, \delta_0) = \mathrm{Inf}_\delta\, r(\xi_0, \delta)$, and our theorem is proved.

3.6 Theorems on Complete Classes of Decision Functions

The notion of admissible decision functions, that of a complete class of decision functions, and that of a minimal complete class were defined in Section 1.3. In this section we shall define some additional notions of completeness and then prove several theorems on complete classes of decision functions.

Let $\mathfrak{D}'$ be a given subset of the set $\mathfrak{D}$ of all decision functions which may be chosen by the experimenter. A class C of decision functions is said to be complete relative to $\mathfrak{D}'$ if for any δ in $\mathfrak{D}'$ not in C we can find a δ^* in C such that δ^* is uniformly better than δ. A class C of decision functions is said to be essentially complete relative to $\mathfrak{D}'$ if for any δ in $\mathfrak{D}'$ we can find an element δ^* of C such that $r(F, \delta^*) \leqq r(F, \delta)$ for all F.[17]

[17] This definition of essential completeness coincides with that given in [66]. In [67], the term "complete class" is used in the sense of an essentially complete class in our present terminology.

In what follows in this section, let $\mathfrak{D}_b$ denote the set of all decision functions δ which are elements of $\mathfrak{D}$ and for which $r(F, \delta)$ is a bounded function of F.

Theorem 3.17. If Assumptions 3.1 to 3.6 hold, the class of all Bayes solutions in the wide sense is complete relative to $\mathfrak{D}_b$.

Proof: Let δ_0 be an element of $\mathfrak{D}_b$ which is not a Bayes solution in the wide sense. Let

$$W^*(F, d^t) = W(F, d^t) - r(F, \delta_0) \tag{3.165}$$

Clearly Assumptions 3.1 to 3.6 remain valid if we replace $W(F, d^t)$ by the weight function $W^*(F, d^t)$. Thus all theorems proved under Assumptions 3.1 to 3.6 can be applied to the decision problem with $W^*(F, d^t)$ as weight function. Let δ_1 be a minimax solution when $W(F, d^t)$ is replaced by $W^*(F, d^t)$. The existence of a minimax solution follows from Theorem 3.7. According to Theorem 3.8, δ_1 is a Bayes solution in the wide sense.[18] Hence, since δ_0 is not a Bayes solution in the wide sense,

$$r(F, \delta_1) \not\equiv r(F, \delta_0) \tag{3.166}$$

for at least one F.

Let $r^*(F, \delta)$ be the risk function when the weight function is given by $W^*(F, d^t)$. Clearly

$$r^*(F, \delta) = r(F, \delta) - r(F, \delta_0) \tag{3.167}$$

Hence $r^*(F, \delta_0) = 0$ identically in F. Since δ_1 is a minimax solution, we must have

$$r^*(F, \delta_1) = r(F, \delta_1) - r(F, \delta_0) \leqq 0 \tag{3.168}$$

for all F. It follows from (3.166) and (3.168) that δ_1 is uniformly better than δ_0, and Theorem 3.17 is proved.

By the closure $\bar{C}$ of a class C of decision functions we shall mean the class of decision functions given as follows: a decision function δ is an element of $\bar{C}$ if and only if δ is either an element of C or a limit (in the regular sense) of a sequence of elements of C.

Theorem 3.18. Let γ be the class of all a priori probability measures ξ for which there exists a finite subset ω of Ω with $\xi(\omega) = 1$. Let C_γ be the class consisting of all decision functions which are Bayes solutions (in the strict sense) relative to members of γ. Then, if Assumptions 3.1 to 3.6 hold, the closure $\bar{C}_\gamma$ of C_γ is essentially complete relative to $\mathfrak{D}_b$.

[18] If a decision function is a Bayes solution in the wide sense when $W^*(F, d^t)$ is the weight function, it retains this property when $W^*(F, d^t)$ is replaced by $W(F, d^t)$, and vice versa.

Proof: Let δ_0 be any element of $\mathfrak{D}_b$, and let $W^*(F, d^t) = W(F, d^t) - W(F, \delta_0)$. As mentioned before, Assumptions 3.1 to 3.6 hold for the decision problem corresponding to the new weight function $W^*(F, d^t)$. Since γ contains the set of all ξ_F as a subset, it follows from Theorem 3.12 that there exists a minimax solution δ_1 of the decision problem corresponding to W^* which is an element of $\overline{C}_\gamma$. Since $r^*(F, \delta_0) = 0$ identically in F, we have

$$r^*(F, \delta_1) = r(F, \delta_1) - r(F, \delta_0) \leqq 0 \tag{3.169}$$

for all F. Hence Theorem 3.18 is proved.

Theorem 3.19. Let ζ be a class of all a priori probability measures ξ with the property that for any ξ not in ζ there exists a sequence $\{\xi_i\}$ $(i = 1, 2, \cdots, ad\ inf.)$ of elements of ζ such that $\lim_{i=\infty} \xi_i(\omega) = \xi(\omega)$ for all subsets ω of Ω. Let C_ζ be the class consisting of all decision functions which are Bayes solutions (in the strict sense) relative to members of ζ. Then, if Assumptions 3.1 to 3.6 hold, the closure $\overline{C}_\zeta$ of C_ζ is essentially complete relative to $\mathfrak{D}_b$.

Proof: Let δ_0 be any element of $\mathfrak{D}_b$ and let $W^*(F, d^t)$ and $r^*(F, \delta)$ be defined as before. If $\lim_{i=\infty} \xi_i(\omega) = \xi_0(\omega)$ for all subsets ω of Ω, then ξ_i converges to ξ_0 in the sense of the metric (3.140) also when $r(\xi, \delta^m)$ is replaced by $r^*(\xi, \delta^m)$. Hence it follows from Theorem 3.12 that there exists a minimax solution δ_1 of the decision problem corresponding to W^* such that δ_1 is a member of $\overline{C}_\zeta$. Since $r^*(F, \delta_0) = 0$ identically in F, $r^*(F, \delta_1) = r(F, \delta_1) - r(F, \delta_0) \leqq 0$ for all F, and Theorem 3.19 is proved.

Theorem 3.20. If Assumptions 3.1 to 3.7 hold, the class of all Bayes solutions in the strict sense is complete relative to $\mathfrak{D}_b$.

This theorem is an immediate consequence of Theorem 3.17 and 3.15.

The classes of decision functions stated in Theorems 3.17 and 3.20 become minimal complete classes if we exclude the Bayes solutions which are not admissible. The conditions under which a Bayes solution is admissible have not yet been thoroughly investigated. The following remarks, however, may be of interest. We shall say that two decision functions δ_1 and δ_2 are equivalent if $r(F, \delta_1) = r(F, \delta_2)$ identically in F. Clearly, if all Bayes solutions relative to a given a priori measure ξ are equivalent, any Bayes solution relative to ξ is admissible. Similarly, if $\{\xi_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ is a given sequence of a priori probability measures, and if all Bayes solutions relative to $\{\xi_i\}$ are equivalent, then any Bayes solution relative to $\{\xi_i\}$ is admissible. A simple sufficient condition for the admissibility of a strict

Bayes solution can be given when the choice of the experimenter is restricted to decision functions δ^m for which the number of stages of experimentation cannot exceed m, where m is a given positive integer. Clearly $r(F, \delta^m)$ is a continuous function of F in the sense of the metric $\rho(F_1, F_2, m)$ given in (3.139). Let ξ be an a priori probability measure such that $\xi(\omega) > 0$ for any subset ω of Ω which is open in the sense of the metric $\rho(F_1, F_2, m)$. Then any Bayes solution relative to ξ must be admissible. Suppose, to the contrary, that δ_1 and δ_2 are Bayes solutions relative to ξ and δ_2 is uniformly better than δ_1. Then there exists an element F_0 of Ω such that $r(F_0, \delta_2) < r(F_0, \delta_1)$. Because of the continuity of $r(F, \delta)$ in F, there exists an open set ω which contains F_0 and is such that $r(F, \delta_2) < r(F, \delta_1)$ for any F in ω. But then $r(\xi, \delta_2) < r(\xi, \delta_1)$ in contradiction to the assumption that both δ_1 and δ_2 are Bayes solutions relative to ξ.

Chapter 4. PROPERTIES OF BAYES SOLUTIONS WHEN THE CHANCE VARIABLES ARE INDEPENDENTLY AND IDENTICALLY DISTRIBUTED AND THE COST OF EXPERIMENTATION IS PROPORTIONAL TO THE NUMBER OF OBSERVATIONS [1]

4.1 Development of the General Theory

4.1.1 Introductory Remarks

In this chapter we shall deal exclusively with the case where the successive chance variables X_1, X_2, $\cdots$, etc., are independently and identically distributed and the cost of experimentation is proportional to the number of observations. Since in this case the cost of experimentation depends only on the total number of observations made, we can restrict ourselves, as was pointed out in Section 3.1.3, to decision functions δ for which each stage of experimentation consists of exactly one observation. Furthermore, because of the assumption that X_1, X_2, $\cdots$, etc., are independently and identically distributed, we can assume without loss of generality that the ith stage of the experiment consists exactly of a single observation on X_i ($i = 1, 2, \cdots$, etc.). Let d_i^e denote the decision to take an observation on X_i. Then our condition on δ can be expressed

$$\delta(D^e \mid 0) = \delta(d_1^e \mid 0) \tag{4.1}$$

$$\delta(D^e \mid x; d_1^e, \cdots, d_i^e) = \delta(d^e{}_{i+1} \mid x; d_1^e, \cdots, d_i^e) \quad (i = 1, 2, \cdots)$$

In what follows in this chapter, we shall restrict ourselves to decision functions δ which satisfy (4.1). We shall use the symbols $\delta(1 \mid 0)$ and $\delta(i + 1 \mid x_1, \cdots, x_i)$ synonymously with $\delta(d_1^e \mid 0)$ and $\delta(d^e{}_{i+1} \mid x; d_1^e, \cdots, d_i^e)$, respectively; i.e.,

$$\delta(1 \mid 0) = \delta(d_1^e \mid 0) \tag{4.2}$$

$$\delta(i + 1 \mid x_1, \cdots, x_i) = \delta(d^e{}_{i+1} \mid x; d_1^e, \cdots, d_i^e)$$

Thus $\delta(1 \mid 0)$ is the probability that we shall take an observation on X_1, and $\delta(i + 1 \mid x_1, \cdots, x_i)$ is the conditional probability that we shall take an observation on X_{i+1} knowing that $X_1, \cdots, X_i$ have been observed and the values $x_1, \cdots, x_i$, respectively, have been obtained.

[1] Most of the results given in this chapter appeared in an earlier publication by Wald and Wolfowitz [71].

For any subset $\bar{D}^t$ of D^t, we shall use the symbol $\delta(\bar{D}^t \mid x_1, \cdots, x_i)$ to denote $\delta(\bar{D}^t \mid x; d_1^e, \cdots, d_i^e)$.

For any element F of Ω, let $f(x_i \mid F)$ denote the elementary probability law of X_i when F is the true distribution of X. Thus $f(x_i \mid F)$ denotes the density function of X_i when F is absolutely continuous, and the probability that $X_i = x_i$ when F is discrete. Let $f^*(x_i \mid F)$ denote the cumulative distribution function of X_i; i.e.,

$$f^*(x_i \mid F) = \int_{-\infty}^{x_i} f(t \mid F)\, dt \tag{4.3}$$

in the absolutely continuous case, and

$$f^*(x_i \mid F) = \sum_{t<x_i} f(t \mid F) \tag{4.4}$$

in the discrete case.

If ξ is the a priori probability measure on Ω, for given values $x_1, \cdots, x_m$ of the first m chance variables the a posteriori probability of a subset ω of Ω is given by

$$\zeta(\omega \mid \xi, x_1, \cdots, x_m) = \frac{\int_\omega f(x_1 \mid F) \cdots f(x_m \mid F)\, d\xi}{\int_\Omega f(x_1 \mid F) \cdots f(x_m \mid F)\, d\xi} \tag{4.5}$$

Let $W(\xi, d^t)$ be defined by

$$W(\xi, d^t) = \int_\Omega W(F, d^t)\, d\xi \tag{4.6}$$

Because of the compactness of the space D^t, $\mathrm{Min}_{d^t} W(\xi, d^t)$ obviously exists.

For any non-negative integer m, let δ^m denote a decision function for which the probability is 1 that the total number of observations will not exceed m. For any a priori probability measure ξ on Ω, we put

$$\begin{aligned} \rho_m(\xi) &= \mathrm{Inf}_{\delta^m} r(\xi, \delta^m) \\ \rho(\xi) &= \mathrm{Inf}_{\delta} r(\xi, \delta) \end{aligned} \tag{4.7}$$

In particular,

$$\rho_0(\xi) = \mathrm{Inf}_{\delta^0} r(\xi, \delta^0) = \mathrm{Min}_{d^t} W(\xi, d^t) \tag{4.8}$$

In the next section we shall study the functions $\rho(\xi)$ and $\rho_m(\xi)$.

It will be assumed throughout this chapter that Assumptions 3.1 to 3.4 hold, even if this is not stated explicitly. Assumption 3.5 is automatically fulfilled, and Assumption 3.6 is replaced, by the assumption that any decision function δ subject to (4.1) may be used.

4.1.2 Properties of the Functions $\rho(\xi)$ and $\rho_m(\xi)$

In this section we shall derive several theorems concerning the functions $\rho(\xi)$ and $\rho_m(\xi)$.

Theorem 4.1. The following recursion formula holds:

$$\rho_{m+1}(\xi) = \text{Min } [\rho_0(\xi), \int_{-\infty}^{\infty} \rho_m(\xi_a)\, df^*(a \mid \xi) + c] \tag{4.9}$$

$$(m = 0, 1, 2, \cdots, \text{ad inf.})$$

where

$$\xi_a(\omega) = \zeta(\omega \mid \xi, a) \tag{4.10}$$

$$f^*(a \mid \xi) = \int_{\Omega} f^*(a \mid F)\, d\xi$$

Proof: Let $\rho_m{}^*(\xi)$ $(m = 1, 2, \cdots, \text{ad inf.})$ denote the infimum of $r(\xi, \delta)$ with respect to δ, where δ is restricted to decision functions for which the probability is 1 that the number of observations is $\geqq 1$ and $\leqq m$. Obviously

$$\rho_{m+1}(\xi) = \text{Min } [\rho_0(\xi), \rho^*{}_{m+1}(\xi)] \tag{4.11}$$

Let $\rho_m{}^*(\xi \mid a)$ denote the infimum with respect to δ of the conditional risk (conditional expected value of $W(F, d^t)$ plus conditional expected value of cost of experimentation) when ξ is the a priori probability measure on Ω, the observed value of X_1 is equal to a, and the choice of the experimenter is restricted to decision functions δ for which the probability is 1 that the number of observations is $\geqq 1$ and $\leqq m$. Clearly

$$\rho^*{}_{m+1}(\xi \mid a) = \rho_m(\xi_a) + c \tag{4.12}$$

Since

$$\rho^*{}_{m+1}(\xi) = \int_{-\infty}^{\infty} \rho^*{}_{m+1}(\xi \mid a)\, df^*(a \mid \xi) \tag{4.13}$$

equation 4.9 follows from (4.11), (4.12), and (4.13).

Theorem 4.2. The function $\rho(\xi)$ satisfies the equation

$$\rho(\xi) = \text{Min } [\rho_0(\xi), \int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a \mid \xi) + c] \tag{4.14}$$

The proof of this theorem is omitted, since it is essentially the same as that of Theorem 4.1.

Theorem 4.3. The following inequalities hold:

$$0 \leqq \rho_m(\xi) - \rho(\xi) \leqq \frac{{W_0}^2}{cm} \quad (m = 1, 2, \cdots, \text{ad inf.}) \tag{4.15}$$

where W_0 denotes the least upper bound of $W(F, d^t)$.

Proof: Let $\{\delta_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ be a sequence of decision functions such that

$$\lim_{i=\infty} r(\xi, \delta_i) = \rho(\xi) \tag{4.16}$$

Also let $P_i(\xi)$ denote the probability that at least m observations will be made when δ_i is the decision function adopted and ξ is the a priori probability measure on Ω. Since $\rho(\xi) \leqq W_0$ and since

$$r(\xi, \delta_i) \geqq cmP_i(\xi) \tag{4.17}$$

it follows from (4.16) that

$$\limsup_{i=\infty} P_i(\xi) \leqq \frac{W_0}{cm} \tag{4.18}$$

Let $\delta_i{}^m$ be the decision function obtained from δ_i as follows: $\delta_i{}^m(1 \mid 0) = \delta_i(1 \mid 0)$, $\delta_i{}^m(r+1 \mid x_1, \cdots, x_r) = \delta_i(r+1 \mid x_1, \cdots, x_r)$ for $r < m$, $\delta_i{}^m(\overline{D}^t \mid 0) = \delta_i(\overline{D}^t \mid 0)$, and $\delta_i{}^m(\overline{D}^t \mid x_1, \cdots, x_r) = \delta_i(\overline{D}^t \mid x_1, \cdots, x_r)$ for any subset $\overline{D}^t$ of D^t and for any $r < m$, and $\delta_i{}^m({d_0}^t \mid x_1, \cdots, x_m) = 1$, where ${d_0}^t$ is a fixed element of D^t. Clearly

$$r(\xi, \delta_i{}^m) \leqq r(\xi, \delta_i) + P_i(\xi)W_0 \tag{4.19}$$

From (4.16), (4.18), and (4.19) it follows that

$$\limsup_{i=\infty} r(\xi, \delta_i{}^m) \leqq \rho(\xi) + \frac{{W_0}^2}{cm} \tag{4.20}$$

Since $\rho_m(\xi)$ cannot exceed the left-hand member of (4.20), the second half of (4.15) follows from (4.20). The first half of (4.15) is obvious, and the proof of our theorem is completed.

An immediate consequence of Theorem 4.3 is the relation [2]

$$\lim_{m=\infty} \rho_m(\xi) = \rho(\xi) \tag{4.21}$$

uniformly in ξ.

A Bayes solution relative to a given a priori probability measure ξ_0 can immediately be given in terms of the functions $\rho(\xi)$ and $\rho_0(\xi)$ as

[2] A proof of (4.21) is contained implicitly in a paper by Arrow, Blackwell, and Girshick [4].

follows: If $\rho(\xi_0) = \rho_0(\xi_0)$, do not take any observation and make a final decision $d_0{}^t$ for which $W(\xi_0, d_0{}^t) = \rho_0(\xi_0)$. If $\rho(\xi_0) < \rho_0(\xi_0)$, take an observation on X_1 and compute the a posteriori probability measure $\xi_{x_1} = \zeta(\omega \mid \xi_0, x_1)$ corresponding to ξ_0 and x_1. If $\rho(\xi_{x_1}) = \rho_0(\xi_{x_1})$, stop experimentation and make a final decision d^t for which $W(\xi_{x_1}, d^t) = \rho_0(\xi_{x_1})$. If $\rho(\xi_{x_1}) < \rho_0(\xi_{x_1})$, take an observation x_2 on X_2. In general, after the observations $x_1, \cdots, x_m$ have been made, take an additional observation if $\rho(\xi_{x_1,\cdots,x_m}) < \rho_0(\xi_{x_1,\cdots,x_m})$, and stop experimentation with a proper terminal decision if $\rho(\xi_{x_1,\cdots,x_m}) = \rho_0(\xi_{x_1,\cdots,x_m})$, where $\xi_{x_1,\cdots,x_m}$ denotes the a posteriori probability measure corresponding to $\xi_0, x_1, \cdots, x_m$.

If the choice of the experimenter is restricted to decision functions δ^m for which the probability is 1 that the total number of observations will not exceed m, the construction of a Bayes solution relative to a given a priori probability measure ξ_0 can easily be carried out with the help of the functions $\rho_0(\xi), \rho_1(\xi), \cdots, \rho_m(\xi)$ as follows: If $\rho_m(\xi_0) = \rho_0(\xi_0)$, a proper terminal decision is made without any experimentation. If $\rho_m(\xi_0) < \rho_0(\xi_0)$, an observation x_1 on X_1 is made and the a posteriori probability measure ξ_{x_1} is determined. If $\rho_{m-1}(\xi_{x_1}) = \rho_0(\xi_{x_1})$, experimentation is stopped with a proper terminal decision. If $\rho_{m-1}(\xi_{x_1}) < \rho_0(\xi_{x_1})$, an observation x_2 on X_2 is made. In general, after the observations $x_1, \cdots, x_k$ have been made ($k \leqq m$), experimentation is continued when $\rho_{m-k}(\xi_{x_1,\cdots,x_k}) < \rho_0(\xi_{x_1,\cdots,x_k})$, and experimentation is stopped with a proper terminal decision when $\rho_{m-k}(\xi_{x_1,\cdots,x_k}) = \rho_0(\xi_{x_1,\cdots,x_k})$. Starting with $\rho_0(\xi)$, the functions $\rho_1(\xi), \cdots, \rho_m(\xi)$ can be determined step by step with the help of the recursion formula (4.9).

A recursive method of construction for Bayes solutions with a fixed finite upper bound m for the number of observations was given also by Arrow, Blackwell, and Girshick [4]. Although their method is applicable also to non-linear cost functions, the number of steps required by their method is of the order m^2 instead of m, owing to the fact that the solution corresponding to $r + 1$ ($r < m$) cannot be obtained in a single step directly from the solution corresponding to r, but through a recursive procedure starting with $m = 1$.

Theorem 4.4. If ξ_1 and ξ_2 are two probability measures on Ω such that[3]

$$\frac{\xi_1(\omega)}{\xi_2(\omega)} \leqq 1 + \epsilon \tag{4.22}$$

for all ω, then

$$\rho(\xi_1) \leqq (1 + \epsilon)\rho(\xi_2) \tag{4.23}$$

[3] The ratio on the left-hand side of (4.22) is defined to be equal to 1 if $\xi_1(\omega) = \xi_2(\omega) = 0$. This remark refers also to any similar ratios that occur later.

Proof: It follows from (4.22) that

$$r(\xi_1, \delta) \leqq r(\xi_2, \delta)(1 + \epsilon) \tag{4.24}$$

for all δ. Hence (4.23) must hold.

This theorem permits the computation of a simple, and in many cases useful, lower bound for the quantity

$$\int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a \mid \xi) \tag{4.25}$$

that occurs in (4.14). A lower bound for (4.25) can be obtained as follows: For any real value a, let ϵ_a be a non-negative value determined so that [4]

$$\frac{\xi(\omega)}{\xi_a(\omega)} \leqq 1 + \epsilon_a \tag{4.26}$$

for all ω. Then

$$\begin{aligned} \int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a \mid \xi) &\geqq \int_{-\infty}^{\infty} \frac{\rho(\xi)}{1 + \epsilon_a}\, df^*(a \mid \xi) \\ &= \rho(\xi) \int_{-\infty}^{\infty} \frac{1}{1 + \epsilon_a}\, df^*(a \mid \xi) \end{aligned} \tag{4.27}$$

Since $\epsilon_a \geqq 0$ and since $\rho_0(\xi) \geqq \rho(\xi)$, we obviously have

$$\begin{aligned} &\rho(\xi) \int_{-\infty}^{\infty} \frac{1}{1 + \epsilon_a}\, df^*(a \mid \xi) \\ &\qquad \geqq \rho(\xi) - [1 - \int_{-\infty}^{\infty} \frac{1}{1 + \epsilon_a}\, df^*(a \mid \xi)]\rho_0(\xi) \end{aligned} \tag{4.28}$$

Hence we obtain the inequality

$$\int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a \mid \xi) \geqq \rho(\xi) - \rho_0(\xi)[1 - \int_{-\infty}^{\infty} \frac{1}{1 + \epsilon_a}\, df^*(a \mid \xi)] \tag{4.29}$$

An upper bound of (4.25) is obtained by replacing ρ by ρ_0; i.e.,

$$\int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a \mid \xi) \leqq \int_{-\infty}^{\infty} \rho_0(\xi_a)\, df^*(a \mid \xi) \tag{4.30}$$

The bounds given in (4.29) and (4.30) may be useful in constructing Bayes solutions, since the following theorem holds.

[4] The improper value ∞ is admitted for ϵ_a.

Theorem 4.5. If

$$\rho_0(\xi) > \int_{-\infty}^{\infty} \rho_0(\xi_a)\, df^*(a \mid \xi) + c \tag{4.31}$$

then $\rho(\xi) < \rho_0(\xi)$. *If*

$$\rho_0(\xi)[1 - \int_{-\infty}^{\infty} \frac{1}{1+\epsilon_a}\, df^*(a \mid \xi)] < c \tag{4.32}$$

then $\rho(\xi) = \rho_0(\xi)$.

This theorem is an immediate consequence of (4.14), (4.29), and (4.30). With the help of this theorem, we can decide whether $\rho(\xi) < \rho_0(\xi)$ or $\rho(\xi) = \rho_0(\xi)$ whenever ξ satisfies (4.31) or (4.32). It is particularly useful when the class of probability measures ξ for which neither (4.31) nor (4.32) holds is small.

The following continuity theorem is an immediate consequence of Theorem 3.6 in Chapter 3.

Theorem 4.6. Let

$$\{\xi_i\} \quad (i = 0, 1, 2, \cdots, \text{ad inf.}) \tag{4.33}$$

be a sequence of probability measures on Ω *such that*

$$\lim_{i=\infty} \xi_i(\omega) = \xi_0(\omega) \tag{4.34}$$

for all subsets ω *of* Ω. *Then*

$$\lim_{i=\infty} \rho(\xi_i) = \rho(\xi_0) \tag{4.35}$$

4.1.3 Characterization of Bayes Solutions

For any probability measures ξ on Ω, one of the following three conditions must hold:

$$\rho_0(\xi) < r(\xi, \delta) \qquad \text{for all } \delta \text{ for which } \delta(1 \mid 0) = 1 \tag{4.36}$$

$$\rho_0(\xi) \leqq r(\xi, \delta) \qquad \text{for all } \delta \text{ for which } \delta(1 \mid 0) = 1 \tag{4.37}$$

$$= r(\xi, \delta) \quad \text{for at least one } \delta \text{ with } \delta(1 \mid 0) = 1$$

$$\rho_0(\xi) > r(\xi, \delta) \quad \text{for at least one } \delta \text{ with } \delta(1 \mid 0) = 1 \tag{4.38}$$

Let $\rho^*(\xi)$ denote the infimum of $r(\xi, \delta)$ with respect to δ, where δ is subject to the restriction $\delta(1 \mid 0) = 1$. It follows from the general existence theorem given in Chapter 3 (Theorem 3.5) that there exists a decision function δ^* such that

$$\rho^*(\xi) = r(\xi, \delta^*) \qquad \text{and} \qquad \delta^*(1 \mid 0) = 1 \tag{4.39}$$

Because of (4.39), the conditions (4.36), (4.37), and (4.38) are equivalent to $\rho_0(\xi) < \rho^*(\xi)$, $\rho_0(\xi) = \rho^*(\xi)$, and $\rho_0(\xi) > \rho^*(\xi)$, respectively.

We shall say that a probability measure ξ on Ω is of the first type if it satisfies (4.36), of the second type if it satisfies (4.37), and of the third type if it satisfies (4.38). Since the a posteriori probability measure defined in (4.5) is also a probability measure on Ω, any a posteriori probability measure will be of one of the above-mentioned three types.

For any sample point x and any decision function δ, let $m(x, \delta)$ denote the smallest non-negative integer with the property that $\delta(m + 1 \mid x_1, \cdots, x_m) = 0$; for $m = 0$, $\delta(m + 1 \mid x_1, \cdots, x_m)$ reduces to $\delta(1 \mid 0)$. We shall now prove the following theorem characterizing Bayes solutions.

Theorem 4.7. A necessary and sufficient condition for a decision function δ_0 to be a Bayes solution relative to a given a priori distribution ξ_0 is that the following three relations be fulfilled for any sample point x (except perhaps on a set whose probability measure is zero when ξ_0 is the a priori probability measure in Ω):

(a) For any $m < m(x,\delta_0)$ the a posteriori probability measure $\zeta(\omega \mid \xi_0, x_1, \cdots, x_m)$ is of either the second or the third type (for $m = 0$, the above a posteriori probability measure reduces to the a priori probability measure ξ_0). If $\zeta(\omega \mid \xi_0, x_1, \cdots, x_m)$ is of the third type, $\delta(m + 1 \mid x_1, \cdots, x_m) = 1$.

(b) For $m = m(x, \delta_0)$, the a posteriori probability measure $\zeta(\omega \mid \xi_0, x_1, \cdots, x_m)$ is of either the first or the second type.

(c) For $m = m(x, \delta_0)$ we have

$$\delta_0(D^t_{x_1, \cdots, x_m} \mid x_1, \cdots, x_m) = 1$$

where $D^t_{x_1, \cdots, x_m}$ denotes the set of all elements d^t of D^t for which

$$W[\zeta(\omega \mid \xi_0, x_1, \cdots, x_m), d^t] = \mathrm{Min}_{d^t} W[\zeta(\omega \mid \xi_0, x_1, \cdots, x_m), d^t]$$

Proof: The sufficiency of (*a*), (*b*), and (*c*) can easily be verified. To prove the necessity of (*a*), (*b*), and (*c*), let us assume that δ_0 is a decision function that violates at least one of the relations (*a*), (*b*), and (*c*) on a set M^* of sample points x whose probability measure $P(M^* \mid \xi_0)$ according to ξ_0 is positive; i.e.,

$$P(M^* \mid \xi_0) = \int_\Omega \left[\int_{M^*} dF(x) \right] d\xi_0 > 0 \tag{4.40}$$

For any δ, the set M^* is Borel measurable, so that the probability (4.40) always exists. The measurability of M^* can be proved with the help of the measurability assumptions in Section 3.1.5 as follows: Let M_1^* be the set of x's for which (*a*) is violated, M_2^* the set of x's

for which (b) is violated, and M_3^* the set of x's for which (c) is violated. It is sufficient to show that $M_i^*(i = 1, 2, 3)$ is measurable. Let M_{it}^* denote the subset of M_i^* for which the first violation of the corresponding condition occurs for the sample $x_1, \cdots, x_t$. We have merely to show that M_{it}^* is measurable for all i and t. The measurability of M_{3t}^* follows from the fact that $m(x, \delta)$ and $\delta(D^t_{x_1, \cdots, x_m} \mid x_1, \cdots, x_m)$ are Borel measurable functions of $x_1, \cdots, x_m$. To show the measurability of M_{1t}^* and M_{2t}^*, it is sufficient to show that the set of samples $x_1, \cdots, x_t$ for which $\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)$ is of type i $(i = 1, 2, 3)$ is measurable. The latter is certainly true if $\rho_0[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)]$ and $\rho^*[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)]$ are Borel measurable functions of $x_1, \cdots, x_t$. On the basis of the measurability assumptions in Section 3.15, the function $\rho_0[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)]$ can easily be seen to be Borel measurable. From (4.9) and (4.21) it follows that $\rho[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)]$ is also a Borel measurable function of $x_1, \cdots, x_t$. The measurability of $\rho^*[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)]$ follows from the relation

$$\rho^*[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t)] = c + \int_{-\infty}^{\infty} \rho[\zeta(\omega \mid \xi_0, x_1, \cdots, x_t, a)]\, df^*[a \mid \zeta(\omega \mid \xi_0, x_1, \cdots, x_t)]$$

Hence M^* is proved to be Borel measurable.

For any x in M^*, let $t(x)$ be the smallest integer $\geqq 0$ such that at least one of the relations (a), (b), and (c) is violated for the finite sample $x_1, x_2, \cdots, x_{t(x)}$. Clearly, if x is a point of M^*, any sample point y for which $y_1 = x_1, \cdots, y_{t(x)} = x_{t(x)}$ is also in M^*. Thus with every sample point x in M^* there is associated a cylindric subset M_x^* of M^* consisting of all points y whose first $t(x)$ coordinates are equal to the corresponding coordinates of x. Clearly M^* can be represented as a sum of such cylindric sets which are disjoint. Let x^0 be a particular point in M^*, and let $M^*_{x^0}$ be the corresponding cylindric subset of M^*. For any decision function δ for which $\delta(i + 1 \mid x_1^0, \cdots, x_i^0) > 0$ for $i = 0, 1, \cdots, t(x^0) - 1$, let $r(\xi_0, \delta, x_1^0, \cdots, x^0_{t(x^0)})$ denote the conditional risk when ξ_0 is the a priori probability measure on Ω, δ is the decision function adopted, and the first $t(x^0)$ observations are equal to $x_1^0, \cdots, x^0_{t(x^0)}$, respectively. In other words, $r(\xi_0, \delta, x_1^0, \cdots, x^0_{t(x^0)})$ is the conditional expected value of the loss $W(F, d^t)$ plus the conditional expected cost of experimentation when ξ_0 is the a priori probability measure in Ω, δ is the decision function adopted, and the observations $x_1^0, \cdots, x^0_{t(x^0)}$ have been made.

We shall now show that there exists a decision function δ_1 such that

$$r(\xi_0, \delta_1) < r(\xi_0, \delta_0) \tag{4.41}$$

We choose the decision function δ_1 such that the following conditions are satisfied: For any x not in M^* we have

$$\delta_1(\overline{D}^t \mid x_1, \cdots, x_i) = \delta_0(\overline{D}^t \mid x_1, \cdots, x_i) \quad (i = 1, 2, \cdots, \text{ad inf.}) \tag{4.42}$$

and

$$\delta_1(i+1 \mid x_1, \cdots, x_i) = \delta_0(i+1 \mid x_1, \cdots, x_i) \quad (i = 1, 2, \cdots, \text{ad inf.}) \tag{4.43}$$

For any x in M^*, δ_1 satisfies the above equations for $i < t(x)$. Furthermore δ_1 satisfies the conditions (a), (b), and (c) of our theorem. Clearly such a decision function δ_1 exists. Let x^0 be a particular point in M^*, and consider the conditional risk

$$r(\xi_0, \delta_1, x_1^0, \cdots, x^0{}_{t(x^0)}) \tag{4.44}$$

Since δ_1 satisfies the conditions (a), (b), and (c), we can easily verify that

$$r(\xi_0, \delta_1, x_1^0, \cdots, x^0{}_{t(x^0)}) = \text{Min}_\delta\, r(\xi_0, \delta, x_1^0, \cdots, x^0{}_{t(x^0)}) \tag{4.45}$$

where δ is restricted to decision functions for which

$$\delta(i+1 \mid x_1^0, \cdots, x_i^0) > 0 \qquad \text{for } i < t(x^0) \tag{4.46}$$

On the other hand, since δ_0 violates one of the conditions (a), (b), (c) for the sample $(x_1^0, \cdots, x^0{}_{t(x^0)})$, we can easily see that

$$r(\xi_0, \delta_0, x_1^0, \cdots, x^0{}_{t(x^0)}) > \text{Min}_\delta\, r(\xi_0, \delta, x_1^0, \cdots, x^0{}_{t(x^0)}) \tag{4.47}$$

where δ is restricted to decision functions which satisfy (4.46). Equation (4.41) follows from (4.40), (4.45), and (4.47). This completes the proof of Theorem 4.7.

A class C of probability measures ξ on Ω will be said to be convex if, for any two elements ξ_1 and ξ_2 of C and for any positive $\lambda < 1$, the probability measure $\xi = \lambda\xi_1 + (1 - \lambda)\xi_2$ is an element of C.

For any element $d_0{}^t$ of D^t, let $C_{i,d_0{}^t}$ denote the class of all probability measures ξ of type i $(i = 1, 2, 3)$ for which

$$W(\xi, d_0{}^t) = \text{Min}_{d^t}\, W(\xi, d^t) \tag{4.48}$$

Let C_{d^t} denote the set theoretical sum of C_{1,d^t} and C_{2,d^t}. We shall now prove the following theorem.

Theorem 4.8. For any element d^t, the classes C_{1,d^t} and C_{d^t} are convex.

Proof: Let ξ_1 and ξ_2 be two elements of C_{1,d^t}. Then for any decision function δ for which $\delta(1 \mid 0) = 1$ we have

$$W(\xi_1, d^t) < r(\xi_1, \delta) \qquad \text{and} \qquad W(\xi_2, d^t) < r(\xi_2, \delta) \tag{4.49}$$

Let $\xi = \lambda\xi_1 + (1 - \lambda)\xi_2$, where λ is a positive number < 1. Clearly

$$W(\xi, d^t) = \lambda W(\xi_1, d^t) + (1 - \lambda)W(\xi_2, d^t) \tag{4.50}$$

and

$$r(\xi, \delta) = \lambda r(\xi_1, \delta) + (1 - \lambda)r(\xi_2, \delta) \tag{4.51}$$

From (4.49), (4.50), and (4.51) we obtain

$$W(\xi, d^t) < r(\xi, \delta) \quad \text{and} \quad W(\xi, d^t) = \operatorname{Min}_{\bar{d}^t} W(\xi, \bar{d}^t) \tag{4.52}$$

Hence ξ is an element of C_{1,d^t}, and the convexity of C_{1,d^t} is proved. The convexity of C_{d^t} can be proved in the same way, replacing $<$ by $\leqq$ in (4.49) and (4.52).

Theorem 4.9. Let ξ_1 be an element of C_{1,d^t} and ξ_2 an element of C_{2,d^t}. Then for any positive $\lambda < 1$ the probability measure $\xi_3 = \lambda\xi_1 + (1 - \lambda)\xi_2$ is an element of C_{1,d^t}.

Proof: Let ξ_1, ξ_2, and ξ_3 be probability measures satisfying the assumptions of Theorem 4.9. Clearly

$$W(\xi_1, d^t) < r(\xi_1, \delta) \quad \text{and} \quad W(\xi_2, d^t) \leqq r(\xi_2, \delta)$$

for any δ for which $\delta(1 \mid 0) = 1$. From this, (4.50), and (4.51) it follows that

$$W(\xi_3, d^t) < r(\xi, \delta)$$

for any δ for which $\delta(1 \mid 0) = 1$. Since

$$W(\xi_3, d^t) = \operatorname{Min}_{\bar{d}^t} W(\xi_3, \bar{d}^t)$$

ξ_3 must be an element of C_{1,d^t}, and our theorem is proved.

We shall say that a set L of probability measures ξ on Ω is a linear manifold, if for any two elements ξ_1 and ξ_2 of L, $\xi = \alpha\xi_1 + (1 - \alpha)\xi_2$ is also an element of L for any real value α for which $\alpha\xi_1 + (1 - \alpha)\xi_2$ is a probability measure. A linear manifold L will be said to be tangent to C_{d^t} if the intersection of L and C_{2,d^t} is not empty, but the intersection of L and C_{1,d^t} is empty.

For any decision function δ and for any element d^t of D^t, let $L(\delta, d^t)$ denote the linear manifold consisting of all probability measures ξ which satisfy the equation

$$W(\xi, d^t) = r(\xi, \delta) \tag{4.53}$$

Theorem 4.10. Let ξ_0 be an element of C_{2,d^t}, and let δ_0 be a decision function such that $\delta_0(1 \mid 0) = 1$ and $W(\xi_0, d^t) = r(\xi_0, \delta_0)$. Then the linear manifold $L(\delta_0, d^t)$ is tangent to C_{d^t}.

Proof: ξ_0 is obviously an element of $L(\delta_0, d^t)$. Thus the intersection of $L(\delta_0, d^t)$ with C_{2,d^t} is not empty. For any element ξ_1 of C_{1,d^t} we have

$$W(\xi_1, d^t) < r(\xi_1, \delta) \tag{4.54}$$

for any δ for which $\delta(1 \mid 0) = 1$. Hence

$$W(\xi_1, d^t) < r(\xi_1, \delta_0) \tag{4.55}$$

and, therefore, ξ_1 cannot be an element of $L(\delta_0, d^t)$. This proves our theorem.

4.1.4 The Case where X_i Can Take Only Two Values

In this section we shall discuss in somewhat more detail the special case where X_i can take only two values, 0 and 1, say. Let ξ_0 be the a priori probability measure, and let ξ_{ij} denote the a posteriori probability measure after i 0's and j 1's have been observed. The probability measure ξ_{00} reduces, of course, to the a priori probability measure ξ_0. Suppose that there exists a positive integer m such that

$$\rho_0(\xi_{mj}) \leqq c \quad \text{and} \quad \rho_0(\xi_{im}) \leqq c \quad (i, j = 0, 1, \cdots, m) \tag{4.56}$$

It follows from (4.56) that

$$\rho(\xi_{mj}) = \rho_0(\xi_{mj}) \quad \text{and} \quad \rho(\xi_{im}) = \rho_0(\xi_{im}) \quad (i, j = 0, 1, 2, \cdots, m) \tag{4.57}$$

Applying formula (4.14) to our special case, we obtain the recursion formula

$$\rho(\xi_{ij}) = \text{Min}\,[\rho_0(\xi_{ij}), p_{ij}\rho(\xi_{i,j+1}) + (1 - p_{ij})\rho(\xi_{i+1,j}) + c] \tag{4.58}$$

where p_{ij} denotes the probability of obtaining the value 1 in a single trial when ξ_{ij} is the a priori probability measure; ie.,

$$p_{ij} = \int_\Omega f(1 \mid F)\, d\xi_{ij} \tag{4.59}$$

With the help of (4.57) and the recursion formula (4.58), the values of $\rho(\xi_{ij})$ can easily be determined step by step for all (i, j) for which $i \leqq m$ and $j \leqq m$. In fact, (4.57) and (4.58) yield the values $\rho(\xi_{m-1,j})$ and $\rho(\xi_{i,m-1})$ for $i \leqq m - 1$ and $j \leqq m - 1$. After the values $\rho(\xi_{m-1,j})$ and $\rho(\xi_{i,m-1})$ have been determined, the recursion formula (4.58) can be used to compute $\rho(\xi_{m-2,j})$ and $\rho(\xi_{i,m-2})$ ($i \leqq m - 2$ and $j \leqq m - 2$), and so on. A Bayes solution can be given in terms of the quantities $\rho(\xi_{ij})$ $(i, j = 0, 1, 2, \cdots, m)$ as follows: If $\rho(\xi_{00}) = \rho_0(\xi_{00})$, a terminal decision d^t is made for which $W(\xi_{00}, d^t) = \rho_0(\xi_{00})$.

If $\rho(\xi_{00}) < \rho_0(\xi_{00})$, experimentation is continued as long as $\rho(\xi_{ij}) < \rho_0(\xi_{ij})$. The first time that $\rho(\xi_{ij}) = \rho_0(\xi_{ij})$ experimentation is stopped with a terminal decision d^t for which $W(\xi_{ij}, d^t) = \rho_0(\xi_{ij})$. It follows from (4.57) that experimentation will stop at some (i, j) for which $i \leqq m$ and $j \leqq m$.

Since X_i can take only the values 0 and 1, any distribution function F of X_i can be represented by a non-negative number $p \leqq 1$, where p denotes the probability that $X_i = 1$. Thus, in the weight function $W(F, d^t)$, we may replace F by p; i.e., $W(p, d^t)$ denotes the loss incurred when p is the true probability that $X_i = 1$ and the terminal decision d^t is made. The space Ω can now be represented by the closed interval [0, 1].

It is of interest to investigate the conditions which guarantee the existence of a positive integer m for which (4.56) holds. In this connection we shall prove the following theorem.

Theorem 4.11. A positive integer m satisfying (4.56) exists if the following three conditions are fulfilled:

(i) *The a priori probability measure ξ_0 assigns a positive probability to any open subset of the interval* [0, 1].

(ii) *If $\lim_{i=\infty} p_i = p_0$, then $\lim_{i=\infty} W(p_i, d^t) = W(p_0, d^t)$ uniformly in d^t.*

(iii) *For any p there exists a terminal decision d^t such that $W(p, d^t) = 0$.*

Proof: Assume that conditions (i), (ii), and (iii) are fulfilled. For any positive ϵ let $P_{ij}(\epsilon)$ denote the a posteriori probability that p lies in the interval $\left(\frac{j}{i+j} - \epsilon, \frac{j}{i+j} + \epsilon\right)$ after i 0's and j 1's have been observed. We can easily verify that, because of (i),

$$\lim_{i=\infty} P_{ij}(\epsilon) = 1 \tag{4.60}$$

uniformly in j, and

$$\lim_{j=\infty} P_{ij}(\epsilon) = 1 \tag{4.61}$$

uniformly in i.

Let $d_{ij}{}^t$ be a terminal decision for which

$$W\left(\frac{j}{i+j}, d_{ij}{}^t\right) = 0 \tag{4.62}$$

The existence of such a terminal decision follows from condition (iii). Also let

$$W_{ij}(\epsilon) = \mathrm{Max}_p\, W(p, d_{ij}{}^t) \tag{4.63}$$

where p is restricted to the interval $\left(\frac{j}{i+j} - \epsilon, \frac{j}{i+j} + \epsilon\right)$. It follows from (4.62) and condition (ii) that

$$\lim_{\epsilon=0} W_{ij}(\epsilon) = 0 \tag{4.64}$$

uniformly in i and j. Clearly

$$W(\xi_{ij}, d_{ij}{}^t) \leqq P_{ij}(\epsilon)W_{ij}(\epsilon) + [1 - P_{ij}(\epsilon)]W_0 \tag{4.65}$$

for any $\epsilon > 0$, where W_0 is an upper bound of $W(p, d^t)$. It follows from (4.60), (4.61), (4.64), and (4.65) that

$$\lim_{i=\infty} W(\xi_{ij}, d_{ij}{}^t) = 0 \tag{4.66}$$

uniformly in j, and

$$\lim_{j=\infty} W(\xi_{ij}, d_{ij}{}^t) = 0 \tag{4.67}$$

uniformly in i. Since $\rho_0(\xi_{ij}) \leqq W(\xi_{ij}, d_{ij}{}^t)$, Theorem 4.11 follows from (4.66) and (4.67).

The bounds for

$$\int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a \mid \xi)$$

given in (4.29) and (4.30) are particularly simple to compute when X_i can take only the values 0 and 1. To illustrate this, consider the case where Ω consists of three points, p_1, p_2, and p_3 (say), and D^t consists of the elements $d_1{}^t$, $d_2{}^t$, and $d_3{}^t$. Let

$$\begin{aligned} W(p_i, d_j{}^t) = W_{ij} &= 1 \quad \text{if } i \neq j \\ &= 0 \quad \text{if } i = j \end{aligned} \tag{4.68}$$

Any probability measure ξ on Ω can be represented by a vector (ξ^1, ξ^2, ξ^3), where ξ^i denotes the probability that p_i is true. Clearly

$$\rho_0(\xi) = 1 - \text{Max}\,(\xi^1, \xi^2, \xi^3) \tag{4.69}$$

Let $\bar{\xi} = (\bar{\xi}^1, \bar{\xi}^2, \bar{\xi}^3)$ denote the a posteriori probability measure when ξ is the a priori probability measure and one trial was performed yielding the value zero. Similarly let $\bar{\bar{\xi}} = (\bar{\bar{\xi}}^1, \bar{\bar{\xi}}^2, \bar{\bar{\xi}}^3)$ denote the a posteriori probability measure when ξ is the a priori probability measure and one trial was performed giving the value 1. Then

$$\bar{\xi}^i = \frac{p_i\xi^i}{\sum_{j=1}^{3} p_j\xi^j} \quad \text{and} \quad \bar{\bar{\xi}}^i = \frac{(1 - p_i)\xi^i}{\sum_{j=1}^{3} (1 - p_j)\xi^j} \tag{4.70}$$

The upper bound given in (4.30) becomes equal to

$$(4.71)\qquad \int_{-\infty}^{\infty} \rho_0(\xi_a)\, df^*(a \mid \xi) = \left(\sum_{i=1}^{3} p_i \xi^i\right) \rho_0(\bar{\xi}) + (1 - \sum_{i=1}^{3} p_i \xi^i) \rho_0(\underline{\xi})$$

Let

$$(4.72)\qquad \mathrm{Max}_{i,j}\left(\frac{p_i}{p_j} - 1\right) \leqq \epsilon \qquad \text{and} \qquad \mathrm{Max}_{i,j}\left(\frac{1 - p_i}{1 - p_j} - 1\right) \leqq \epsilon$$

Then we can put $\epsilon_a = \epsilon$, and the lower bound given in (4.29) becomes equal to

$$(4.73)\qquad \rho(\xi) - \rho_0(\xi)\left(1 - \frac{1}{1 + \epsilon}\right) = \rho(\xi) - \rho_0(\xi)\frac{\epsilon}{1 + \epsilon}$$

Applying Theorem 4.5, we arrive at the following result: If

$$(4.74)\qquad \rho_0(\xi) > \left(\sum_{i=1}^{3} p_i \xi^i\right) \rho_0(\bar{\xi}) + (1 - \sum_{i=1}^{3} p_i \xi^i) \rho_0(\underline{\xi}) + c$$

then $\rho(\xi) < \rho_0(\xi)$. If

$$(4.75)\qquad \rho_0(\xi)\frac{\epsilon}{1 + \epsilon} < c$$

then $\rho(\xi) = \rho_0(\xi)$.

An example.[5] The method given in this section has been applied to obtain a Bayes solution for the following problem: Let $X_1, X_2, \cdots$, etc., be independently and identically distributed chance variables. The chance variable X_i can take only the values 0 and 1; let p denote the probability that $X_i = 1$. The value p is unknown, and the problem is to test the hypothesis H that $p < \frac{1}{2}$. Let $d_1{}^t$ denote the decision to accept H, and $d_2{}^t$ the decision to reject H. We assume that D^t consists of the elements $d_1{}^t$ and $d_2{}^t$, and that

$$W(p, d_1{}^t) = 0 \quad \text{for } p < \tfrac{3}{4}, \qquad = 1 \quad \text{for } p \geqq \tfrac{3}{4}$$

$$W(p, d_2{}^t) = 0 \quad \text{for } p > \tfrac{1}{4}, \qquad = 1 \quad \text{for } p \leqq \tfrac{1}{4}$$

The cost of experimentation is assumed to be proportional to the number of observations. Let $c = 0.004$ be the cost of a single observation. Furthermore let the a priori distribution of p be the rectangular distribution with the range [0, 1]. In Table I the numbers in the upper halves of the cells give the values of $\rho_0(\xi_{ij})$ for $i, j = 0, 1, \cdots, 10$ and the value $\rho_0(\xi_{10,11}) = \rho_0(\xi_{11,10})$. Since $\rho_0(\xi_{10,j}) = \rho_0(\xi_{j,10}) < c$ for $j < 10$, we see that

$$\rho(\xi_{10,j}) = \rho_0(\xi_{10,j}) \qquad \text{and} \qquad \rho(\xi_{j,10}) = \rho_0(\xi_{j,10})$$

[5] The author is indebted to Mr. Milton Sobel for carrying out the computations for this example.

TABLE I

Number of 0's	Number of 1's: 0	1	2	3	4	5	6	7	8	9	10	11
0	.2500	.0625	.0156	.0039	.0010	.0002	.0001	.0000	.0000	.0000	.0000	
	.0252	.0212	.0126	.0039	.0010	.0002	.0001	.0000	.0000	.0000	.0000	
1	.0625	.1563	.0508	.0156	.0046	.0013	.0004	.0001	.0000	.0000	.0000	
	.0212	.0265	.0225	.0141	.0046	.0013	.0004	.0001	.0000	.0000	.0000	
2	.0156	.0508	.1035	.0376	.0129	.0042	.0013	.0005	.0001	.0000	.0000	
	.0126	.0225	.0250	.0210	.0129	.0042	.0013	.0005	.0001	.0000	.0000	
3	.0039	.0156	.0376	.0706	.0273	.0100	.0035	.0012	.0004	.0001	.0000	
	.0039	.0141	.0210	.0225	.0185	.0100	.0035	.0012	.0004	.0001	.0000	
4	.0010	.0046	.0129	.0273	.0489	.0198	.0076	.0028	.0010	.0003	.0001	
	.0010	.0046	.0129	.0185	.0210	.0161	.0076	.0028	.0010	.0003	.0001	
5	.0002	.0013	.0042	.0100	.0198	.0343	.0143	.0056	.0022	.0008	.0003	....
	.0002	.0013	.0042	.0100	.0161	.0176	.0136	.0056	.0022	.0008	.0003	....
6	.0001	.0004	.0013	.0035	.0076	.0143	.0243	.0103	.0042	.0016	.0006	
	.0001	.0004	.0013	.0035	.0076	.0136	.0143	.0103	.0042	.0016	.0006	
7	.0000	.0001	.0005	.0012	.0028	.0056	.0103	.0173	.0075	.0031	.0012	....
	.0000	.0001	.0005	.0012	.0028	.0056	.0103	.0115	.0075	.0031	.0012	....
8	.0000	.0000	.0001	.0004	.0010	.0022	.0042	.0075	.0124	.0054	.0023	
	.0000	.0000	.0001	.0004	.0010	.0022	.0042	.0075	.0094	.0054	.0023	
9.	.0000	.0000	.0000	.0001	.0003	.0008	.0016	.0031	.0054	.0090	.0039	....
	.0000	.0000	.0000	.0001	.0003	.0008	.0016	.0031	.0054	.0079	.0039	....
10	.0000	.0000	.0000	.0000	.0001	.0003	.0006	.0012	.0023	.0039	.0064	.0028
	.0000	.0000	.0000	.0000	.0001	.0003	.0006	.0012	.0023	.0039	.0074	....
11										....	.0028	
										....	....	

for $j < 10$. Using the recursion formula (4.58) for $i = j = 10$ and the given value of $\rho_0(\xi_{10,11}) = \rho_0(\xi_{11,\,10})$, we find that

$$\rho(\xi_{10,10}) = \rho_0(\xi_{10,10})$$

The values of $\rho(\xi_{ij})$ for $i, j = 0, 1, \cdots, 10$, as given in the lower halves of the cells in Table I, were obtained step by step by repeated application of the recursion formula (4.58).

The heavy lines in Table I include those cells (i, j) for which $\rho(\xi_{ij}) < \rho_0(\xi_{ij})$. In all other cells (i, j) we have $\rho(\xi_{ij}) = \rho_0(\xi_{ij})$. Thus a Bayes solution is given by the following rule: Continue taking observations as long as the pair (i, j) is represented by a cell inside the heavy lines, where i is the number of 0's and j is the number of 1's obtained. At the first time when (i, j) is represented by a cell outside the heavy

line, stop experimentation. At the termination of experimentation, accept H if $i > j$, reject H if $i < j$; either decision can be made if $i = j$.

4.2 Application of the General Theory to the Case where Ω and D^t Are Finite

4.2.1 The Case where Ω Consists of Two Elements

In this section we shall apply the general results of Section 4.1 to the special case where Ω consists of two elements F_1 and F_2 (say), and D^t consists of two elements ${d_1}^t$ and ${d_2}^t$. Here ${d_i}^t$ denotes the terminal decision to accept the hypothesis H_i that F_i is true ($i = 1, 2$). Let

$$W(F_i, {d_j}^t) = W_{ij} = 0 \quad \text{for } i = j \qquad \text{and} \qquad > 0 \quad \text{for } i \neq j \tag{4.76}$$

An a priori probability measure is now given by a vector $\xi = (\xi^1, \xi^2)$, where the component ξ^i is the probability that the true distribution F is equal to $F_i (i = 1, 2)$. Of course, $\xi^i \geqq 0$ and $\xi^1 + \xi^2 = 1$.

Let ξ_1 be the a priori probability distribution given by the vector (1,0), and ξ_2 the a priori probability distribution given by the vector (0, 1). Clearly $C_{{d_1}^t}$ contains ξ_1 but not ξ_2, and $C_{{d_2}^t}$ contains ξ_2 but not ξ_1. It follows from Theorems 4.6 and 4.8 that $C_{{d_1}^t}$ and $C_{{d_2}^t}$ are closed and convex sets of probability measures ξ. Furthermore we obviously have

$$\xi^2 W_{21} \leqq \xi^1 W_{12} \tag{4.77}$$

for all ξ in $C_{{d_1}^t}$, and

$$\xi^2 W_{21} \geqq \xi^1 W_{12} \tag{4.78}$$

for all ξ in $C_{{d_2}^t}$. Let $\xi_0 = ({\xi_0}^1, {\xi_0}^2)$ be the probability measure for which

$${\xi_0}^2 W_{21} = {\xi_0}^1 W_{12} \tag{4.79}$$

Clearly, because of (4.77) and (4.78), $\xi^2 \leqq {\xi_0}^2$ for any ξ in $C_{{d_1}^t}$ and $\xi^2 \geqq {\xi_0}^2$ for any ξ in $C_{{d_2}^t}$. Since $C_{{d_1}^t}$ and $C_{{d_2}^t}$ are closed and convex, there exist two positive numbers h' and h'' such that

$$0 < h' \leqq {\xi_0}^2 \leqq h'' < 1 \tag{4.80}$$

and such that the class $C_{{d_1}^t}$ consists of all ξ for which $\xi^2 \leqq h'$, and the class $C_{{d_2}^t}$ consists of all ξ for which $\xi^2 \geqq h''$. It follows from Theorem 4.9 that $C_{2,{d_1}^t}$ consists of the single element ξ for which $\xi^2 = h'$, and $C_{2,{d_2}^t}$ consists of the single element ξ for which $\xi^2 = h''$.

Applying Theorem 4.7, we arrive at the following characterization of a Bayes solution corresponding to a given a priori probability measure: *Let ξ_i denote the a posteriori probability measure after i observations have been made ($i = 1, 2, \cdots$, ad inf.), and let ξ_0 be the a priori prob-*

ability measure. If $\xi_0^2 < h'$, *accept* H_1. *If* $\xi_0^2 = h'$, *decide between accepting* H_1 *and taking an observation on* X_1 *by any independent chance mechanism. If* $h' < \xi_0^2 < h''$, *take an observation on* X_1. *If* $\xi_0^2 = h''$, *decide between accepting* H_2 *and taking an observation on* X_1 *by any independent chance mechanism. If* $\xi_0^2 > h''$, *accept* H_2. *If the foregoing procedure resulted in taking an observation on* X_1, *compute* ξ_1 *and proceed in a similar way, except that* ξ_0 *is now replaced by* ξ_1 *and* X_1 *by* X_2. *If this rule results in taking an observation on* X_2, *compute* ξ_2, *and so on.*

The a posteriori probability ξ_i^2 of H_2 after i observations have been made is given by

$$\xi_i^2 = \frac{\xi_0^2 f(x_1 \mid F_2) \cdots f(x_i \mid F_2)}{\xi_0^1 f(x_1 \mid F_1) \cdots f(x_i \mid F_1) + \xi_0^2 f(x_1 \mid F_2) \cdots f(x_i \mid F_2)} \tag{4.81}$$

The Bayes solution described above is identical to the sequential probability ratio test procedure for deciding between H_1 and H_2. The sequential probability ratio test is defined as follows (see [65]): For any positive integer i, let

$$\frac{p_{2i}}{p_{1i}} = \frac{f(x_1 \mid F_2) \cdots f(x_i \mid F_2)}{f(x_1 \mid F_1) \cdots f(x_i \mid F_1)} \tag{4.82}$$

and let

$$\frac{p_{20}}{p_{10}} = 1 \tag{4.83}$$

Two positive constants A and B ($B \leqq A$) are chosen. The procedure for carrying out the experimentation and making a terminal decision is identical to the procedure for the above-described Bayes solution, except that ξ_i^2 is replaced by p_{2i}/p_{1i}, h' by B, and h'' by A. Since p_{2i}/p_{1i} is a strictly monotonic function of ξ_i^2, the above-described Bayes solution coincides with the sequential probability ratio test for properly chosen values of the constants A and B.

The above definition of a sequential probability ratio test differs slightly from the one given in earlier publications (see [65]). In [65] the constant B is restricted to values <1 and the constant A to values >1. No such restriction is made here. Furthermore in [65] it is required that experimentation be stopped when $p_{2m}/p_{1m} = A$ or $= B$, whereas here experimentation may be continued in this case. Of course, if the probability (under H_1 and H_2) is zero that $p_{2m}/p_{1m} = A$ or $= B$, as it usually is when the distributions are absolutely continuous, the difference between the two definitions of the sequential probability ratio test is of no consequence.

It follows from Theorem 3.20 in Chapter 3 that the class of all Bayes solutions for deciding between H_1 and H_2 is a complete class. Since any Bayes solution is equivalent to a sequential probability ratio test corresponding to some values of the constants A and B, we arrive at the following result.

Theorem 4.12. The class of all sequential probability ratio tests corresponding to all possible values of the constants A and B is a complete class of decision functions for deciding between H_1 and H_2.[6]

4.2.2 The Case where Ω Contains More than Two Elements

It will be sufficient to discuss the case when Ω consists of three elements F_1, F_2, and F_3, since the extension to any finite number >3 will be obvious. Let

$$W(F_i, d_j^t) = W_{ij} = 0 \quad \text{for } i = j \qquad (i, j = 1, 2, 3)$$
$$> 0 \quad \text{for } i \neq j$$

Any a priori distribution $\xi = (\xi^1, \xi^2, \xi^3)$ can be represented by a point with the coordinates ξ^1, ξ^2, and ξ^3. The totality of all possible a priori distributions ξ will fill out the triangle T with the vertices V_1, V_2, V_3, where V_i represents the a priori distribution ξ whose ith component ξ^i is equal to 1 (see Fig. 1). Clearly the vertex V_i is contained in $C_{d_i^t}$. Thus, according to Theorem 4.8, the set $C_{d_i^t}$ ($i = 1, 2, 3$) is a convex subset of T containing V_i.

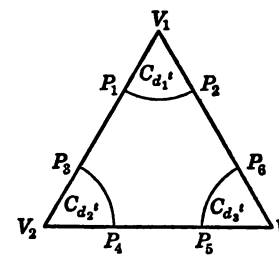

FIG. 1

If one of the components of ξ, say ξ^i, is zero, then H_i can be disregarded, and the problem of constructing Bayes solutions reduces to the previously considered case where $k = 2$. Thus, in particular, the determination of the boundary points $P_1, P_2, \cdots, P_6$ of $C_{d_1^t}$, $C_{d_2^t}$, and $C_{d_3^t}$, which are on the boundary of T, reduces to the previously discussed case where $k = 2$.

We shall now show that C_{2,d_i^t} consists precisely of the boundary points of $C_{d_i^t}$, provided that the sets $C_{d_1^t}$, $C_{d_2^t}$, and $C_{d_3^t}$ are disjoint. It follows immediately from Theorem 4.9 that C_{2,d_i^t} must be a subset of the boundary of $C_{d_i^t}$. Thus we have merely to show that if ξ_0 is a boundary point of $C_{d_i^t}$, then ξ_0 is a point of C_{2,d_i^t}. Since ξ_0 is a point of

[6] This theorem follows also from an optimum property of the sequential probability ratio test proved by Wald and Wolfowitz [69].

$C_{d_i t}$, we must have $\rho_0(\xi_0) \leqq \rho^*(\xi_0)$. Let $\{\xi_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$ be a sequence of a priori measures such that $\lim_{i=\infty} \xi_i = \xi_0$ and ξ_i $(i > 0)$ is not contained in any of the sets $C_{d_1 t}$, $C_{d_2 t}$, and $C_{d_3 t}$. Such a sequence $\{\xi_i\}$ exists, since ξ_0 is a boundary point of $C_{d_i t}$ and the sets $C_{d_1 t}$, $C_{d_2 t}$, $C_{d_3 t}$ are disjoint. Clearly $\rho_0(\xi_i) > \rho^*(\xi_i)$ for $i = 1, 2, \cdots$, ad inf. Since $\lim_{i=\infty} \rho_0(\xi_i) = \rho_0(\xi_0)$ and $\lim_{i=\infty} \rho^*(\xi_i) = \rho^*(\xi_0)$ [the continuity of $\rho^*(\xi)$ can be proved in the same way as that of $\rho(\xi)$], we must have $\rho_0(\xi_0) \geqq \rho^*(\xi_0)$. Hence $\rho_0(\xi_0) = \rho^*(\xi_0)$ and, therefore, ξ_0 must be an element of $C_{2,d_i t}$.

Tangents to the sets $C_{d_1 t}$, $C_{d_2 t}$, and $C_{d_3 t}$ can be constructed at the boundary points $P_1, P_2, \cdots, P_6$ as follows: Consider, for example, the boundary point P_1 of $C_{d_1 t}$ (Fig. 1) which is on the line $V_1 V_2$. Let ξ_1 be the probability distribution represented by the point P_1. Since the a priori probability of H_3 is zero according to ξ_1, we can disregard H_3 in constructing a Bayes solution relative to ξ_1. Let δ_1 be a sequential probability ratio test for testing H_1 against H_2 such that δ_1 is a Bayes solution relative to ξ_1 and $\delta_1(1 \mid 0) = 1$. Since ξ_1 is a boundary point, such a decision function δ_1 exists. Thus we have

$$W(\xi_1, d_1{}^t) = r(\xi_1, \delta_1) = \operatorname{Inf}_\delta r(\xi_1, \delta) \tag{4.84}$$

Let α_{ij} denote the probability of accepting H_j when H_i is true and δ_1 is the decision function adopted. Also let n_i denote the expected number of observations when H_i is true and δ_1 is adopted. Then for any a priori probability measure ξ we have

$$r(\xi, \delta_1) = \sum_{i,j} \xi^i W_{ij} \alpha_{ij} + c \sum_i \xi^i n_i \tag{4.85}$$

and

$$W(\xi, d_1{}^t) = \sum_i \xi^i W_{i1} \tag{4.86}$$

Thus the linear manifold $L(\delta_1, d_1{}^t)$ is simply the straight line given by the equation

$$\sum_i \xi^i W_{i1} = \sum_{i,j} \xi^i W_{ij} \alpha_{ij} + c \sum_i \xi^i n_i \tag{4.87}$$

This straight line goes through P_1 and, because of Theorem 4.10, it is tangent to $C_{d_1 t}$. Tangents at the points $P_2, P_3, \cdots, P_6$ can be constructed in a similar way.

More general results concerning the case of finite Ω and D^t, admitting also non-linear cost functions, were obtained by Arrow, Blackwell, and Girshick [4].

Chapter 5. APPLICATION OF THE GENERAL THEORY TO VARIOUS SPECIAL CASES

5.1 Discussion of Some Non-Sequential Decision Problems

5.1.1 Non-Sequential Decision Problems when the Spaces Ω and D^t Are Finite

By non-sequential decision functions we mean decision functions δ according to which the probability is 1 that experimentation is carried out in a single stage. In this section we shall discuss decision problems for which the spaces Ω and D^t are finite and experimentation is carried out in a single stage by observing the values of the first N chance variables in the sequence $\{X_i\}$ $(i = 1, 2, \cdots, \text{ad inf.})$. Let $F_1, \cdots, F_k$ be the elements of Ω, and $d_1{}^t, \cdots, d_u{}^t$ the elements of D^t. Since experimentation is carried out in a single stage by observing the values of $X_1, \cdots, X_N$, any decision function δ can be represented by a vector function $\delta(x_1, \cdots, x_N)$ with u components $\delta_1(x_1, \cdots, x_N), \cdots, \delta_u(x_1, \cdots, x_N)$ satisfying the conditions

$$\delta_i(x_1, \cdots, x_N) \geqq 0 \qquad (i = 1, \cdots, u)$$

$$\sum_{i=1}^{u} \delta_i(x_1, \cdots, x_N) = 1 \tag{5.1}$$

Here x_i denotes the observed value of X_i, and $\delta_i(x_1, \cdots, x_N)$ is the probability that we shall make the terminal decision $d_i{}^t$ when the sample $(x_1, \cdots, x_N)$ is observed. After the sample $(x_1, \cdots, x_N)$ has been obtained, the actual selection of the terminal decision d^t is made with the help of a chance mechanism constructed in such a way that the probability that $d_i{}^t$ will be selected is equal to $\delta_i(x_1, \cdots, x_N)$. In the special case when the functions $\delta_i(x_1, \cdots, x_N)$ $(i = 1, \cdots, u)$ can take only the values 0 and 1, we have a non-randomized decision function.

Let $f_i(x_1, \cdots, x_N)$ denote the joint elementary probability law of $X_1, \cdots, X_N$ when F_i is the true distribution, i.e., $f_i(x_1, \cdots, x_N)$ denotes the probability density at $x_1, \cdots, x_N$ when F_i is absolutely continuous, and the probability that $X_j = x_j$ for all values $j \leqq N$ when F_i is discrete. Let W_{ij} denote the loss $W(F_i, d_j{}^t)$ when F_i is the true distribution and the terminal decision $d_j{}^t$ is adopted. Since only decision functions are admitted for which experimentation is carried

out in a single stage by observing the values of $X_1, \cdots, X_N$, the cost of experimentation does not depend on the choice of the decision function and, therefore, it can be disregarded altogether. The risk when F_i is true and the decision function δ is adopted is then given by

$$(5.2) \qquad r(F_i, \delta) = \sum_{j=1}^{u} \int_{M_N} W_{ij}\, \delta_j(x_1, \cdots, x_N)\, dF_i(x_1, \cdots, x_N)$$

where M_N denotes the space of all samples $(x_1, \cdots, x_N)$, and $F_i(x_1, \cdots, x_N)$ denotes the joint cumulative distribution of $X_1, \cdots, X_N$.

We shall now study the nature of the Bayes solutions of the decision problem. Any a priori distribution in Ω can be represented by a vector $\xi = (\xi_1, \cdots, \xi_k)$, where ξ_i denotes the a priori probability that F_i is true. After the sample $(x_1, \cdots, x_N)$ has been drawn, the a posteriori probability that F_i is true is given by

$$(5.3) \qquad \xi_i^* = \frac{\xi_i f_i(x_1, \cdots, x_N)}{\sum_{j=1}^{k} \xi_j f_j(x_1, \cdots, x_N)} \qquad (i = 1, 2, \cdots, k)$$

The a posteriori risk associated with the terminal decision d_j^t, i.e., the a posteriori expected value of $W(F, d_j^t)$ (determined on the basis of the a posteriori distribution in Ω), is given by

$$(5.4) \qquad r_j(x_1, \cdots, x_N) = \sum_{i=1}^{k} \xi_i^* W_{ij} \qquad (j = 1, 2, \cdots, u)$$

The following characterization theorem holds.

Theorem 5.1. A necessary and sufficient condition for a decision function δ to be a Bayes solution relative to a given a priori probability measure ξ is that

$$(5.5) \qquad \delta_j(x_1, \cdots, x_N) = 0$$

for any sample $(x_1, \cdots, x_N)$ (except perhaps on a set of ξ-measure zero)[1] *and for any j for which*

$$(5.6) \qquad r_j(x_1, \cdots, x_N) > \text{Min}\,[r_1(x_1, \cdots, x_N), \cdots, r_u(x_1, \cdots, x_N)]$$

The proof of this theorem is very simple and is omitted. Since $r_j(x_1, \cdots, x_N)$ is proportional to the function

$$(5.7) \qquad t_j(x_1, \cdots, x_N) = \sum_{i=1}^{k} \xi_i W_{ij} f_i(x_1, \cdots, x_N)$$

we can replace r_j by t_j in the above theorem.

[1] By the ξ-measure of a subset R of the sample space we mean $\sum_{i=1}^{k} \xi_i \int_R dF_i$.

If for any pair i, j the set of all samples $(x_1, \cdots, x_N)$ for which $t_i = t_j$ is of ξ-measure zero, then Theorem 5.1 shows that the following Bayes solution is essentially unique: Take the terminal decision d_j^t, where j is the smallest positive integer satisfying the equation $t_j = \text{Min}\ (t_1, \cdots, t_u)$. Any other Bayes solution can differ from this particular one only on a set of ξ-measure zero.

Applying Theorem 3.20 of Chapter 3, we obtain the following theorem.

Theorem 5.2. The class of all Bayes solutions δ corresponding to all possible a priori probability measures ξ is a complete class of decision functions.

In what follows in this and subsequent sections of the present chapter, we shall regard two decision functions δ^1 and δ^2 as identical if δ^1 differs from δ^2 only on a set of sample points x whose probability measure is zero under any element F of Ω.

If for any a priori probability measure there exists only one Bayes solution, the class of all Bayes solutions is merely a $(k-1)$-parameter family of decision functions.

Theorems 3.7, 3.9, 3.10, and 3.14 of Chapter 3 yield immediately the following theorem.

Theorem 5.3. There exist an a priori distribution $\xi^0 = (\xi_1{}^0, \cdots, \xi_k{}^0)$ and a decision function δ^0 such that

(i) *δ^0 is a Bayes solution relative to ξ^0.*

(ii) *δ^0 is a minimax solution, i.e., $Max_i\ r(F_i, \delta^0) \leqq Max_i\ r(F_i, \delta)$ for all δ.*

(iii) *For any i for which $\xi_i{}^0 > 0$, we have $r(F_i, \delta^0) = Max_j\ r(F_j, \delta^0)$.*

(iv) *ξ^0 is a least favorable a priori distribution; i.e.,*

$$\text{Inf}_\delta \left[\sum_{i=1}^{k} \xi_i{}^0 r(F_i, \delta) \right] \geqq \text{Inf}_\delta \left[\sum_{i=1}^{k} \xi_i r(F_i, \delta) \right]$$

for any ξ.

Because of Theorem 3.9, the essential difficulty in constructing a minimax solution is solved if we can find a least favorable a priori distribution ξ^0. We have merely to study the Bayes solutions relative to ξ^0, at least one of which must be a minimax solution. A Bayes solution δ^0 relative to ξ^0 will be a minimax solution if and only if

$$r(F_i, \delta^0) = \text{Max}_j\ r(F_j, \delta^0) \tag{5.8}$$

for all i for which $\xi_i{}^0 > 0$.

As to the problem of finding a least favorable a priori distribution ξ^0, the following remarks may be helpful. For any a priori distribution ξ, let δ_ξ be the particular Bayes solution given by the following rule: Decide on d_j^t where j is the smallest integer for which

$$t_j(x_1, \cdots, x_N) = \text{Min } [t_1(x_1, \cdots, x_N), \cdots, t_u(x_1, \cdots, x_N)] \tag{5.9}$$

Consider the average risk

$$r(\xi, \delta_\xi) = \sum_{i=1}^{k} \xi_i r(F_i, \delta_\xi) \tag{5.10}$$

This is a function of $\xi_1, \xi_2, \cdots, \xi_k$ only. An a priori distribution ξ^0 is a least favorable one, if it maximizes $r(\xi, \delta_\xi)$; i.e., if

$$r(\xi^0, \delta_{\xi^0}) \geqq r(\xi, \delta_\xi) \tag{5.11}$$

for all ξ. Thus the problem of finding a least favorable distribution is reduced to the problem of finding a probability measure ξ^0 for which (5.11) is satisfied.

We shall now apply Theorems 5.1, 5.2, and 5.3 to the case where the number of elements in Ω, as well as in D^t, is equal to two, and $d_j^t (j = 1, 2)$ represents the decision to accept the hypothesis that F_j is the true distribution. Since accepting the hypothesis that F_i is true when F_i is actually true is a correct decision, we put $W_{11} = W_{22} = 0$, while W_{12} and W_{21} are assumed to be positive. Then Theorem 5.1 gives the following necessary and sufficient condition for a decision function δ to be a Bayes solution relative to an a priori distribution ξ: $\delta_1(x_1, \cdots, x_N) = 0$ whenever $\xi_1 W_{12} f_1(x_1, \cdots, x_N) < \xi_2 W_{21} f_2(x_1, \cdots, x_N)$, and $\delta_1(x_1, \cdots, x_N) = 1$ whenever $\xi_1 W_{12} f_1(x_1, \cdots, x_N) > \xi_2 W_{21} f_2(x_1, \cdots, x_N)$ (except perhaps on a set of ξ-measure zero). When $\xi_1 W_{12} f_1(x_1, \cdots, x_N) = \xi_2 W_{21} f_2(x_1, \cdots, x_N)$, $\delta_1(x_1, \cdots, x_N)$ may take any value in the closed interval [0, 1].

We shall say that δ is a probability ratio decision function if there exists a non-negative constant h (the value $h = \infty$ is also admitted) such that

$$\delta_1(x_1, \cdots, x_N) = 0 \qquad \text{whenever} \frac{f_2(x_1, \cdots, x_N)}{f_1(x_1, \cdots, x_N)} > h$$

and

$$\delta_1(x_1, \cdots, x_N) = 1 \qquad \text{whenever} \frac{f_2(x_1, \cdots, x_N)}{f_1(x_1, \cdots, x_N)} < h$$

Clearly any Bayes solution is a probability ratio decision function (except perhaps on a set of ξ-measure zero). We can easily verify that the converse is also true; i.e., if δ is a probability ratio decision

function, there exists an a priori distribution ξ such that δ is a Bayes solution relative to ξ. Thus the class of all Bayes solutions coincides with the class of all probability ratio decision functions. Hence the class of all probability ratio decision functions is a complete class.

The above results are closely related to a well-known theorem by Neyman and Pearson [37]. They have shown that, if δ is a probability ratio decision function corresponding to some finite and positive value h, δ is an admissible decision function; i.e., no uniformly better decision function exists. This theorem follows immediately from the fact that a probability ratio decision function corresponding to a finite and positive value h is a Bayes solution relative to some ξ with positive components. The complete class theorem is, in a sense, the converse of the Neyman-Pearson theorem. While the Neyman-Pearson theorem shows that for a decision function δ to be admissible it is sufficient that δ be a probability ratio decision function, the complete class theorem shows that this is also necessary.

The special case where the number of elements in D^t is equal to that in Ω and where

$$W_{ij} = 1 \quad \text{for } i \neq j \tag{5.12}$$
$$= 0 \quad \text{for } i = j$$

is of particular interest. In this case we may interpret d_i^t as the decision to accept the hypothesis that F_i is the true distribution. The risk $r(F_i, \delta)$ is then simply the probability of making a wrong decision when F_i is true and δ is the decision function adopted.

Theorem 5.4. All components of a least favorable distribution ξ must be positive if the following conditions are fulfilled:

(i) *The number of elements in D^t is equal to the number of elements in Ω.*

(ii) *The quantities W_{ij} satisfy (5.12).*

(iii) *There exists a decision function δ such that $r(F_i, \delta) < 1$ for all i.*

(iv) *If R is a subset of M_N for which, $\int_R dF_i = 0$ for some i, then $\int_R dF_i = 0$ for all values of i.*

Proof: Let k be the number of elements in Ω and let δ^0 be a minimax solution. Since, by assumption, there exists a decision function δ for which $r(F_i, \delta) < 1$ for all i, we have

$$r(F_i, \delta^0) < 1 \tag{5.13}$$

for $i = 1, 2, \cdots, k$. Let ξ^0 be a least favorable a priori probability measure. Then δ^0 is a Bayes solution relative to ξ^0. Suppose that

one of the components of ξ^0, say $\xi_1{}^0$, is zero. It follows from Theorem 5.1 that for any decision function $\delta = [\delta_1(x_1, \cdots, x_N), \cdots, \delta_k(x_1, \cdots, x_N)]$, which is a Bayes solution relative to ξ^0, we must have $\delta_1(x_1, \cdots, x_N) = 0$, except perhaps on a set whose ξ^0-measure is zero. Thus, in particular, $\delta_1{}^0(x_1, \cdots, x_N) = 0$, except perhaps on a set of ξ^0-measure zero. Because of condition (iv) of our theorem, the exceptional set R in which $\delta_1{}^0(x_1, \cdots, x_N) \neq 0$ must satisfy the equation $\int_R dF_i = 0$ for $i = 1, 2, \cdots, k$. Hence $r(F_1, \delta^0) = 1$. But this contradicts (5.13), and our theorem is proved.

If Ω and D^t have the same number of elements and if (5.12) holds, a Bayes solution relative to a given a priori probability measure ξ is given by the following simple rule: Decide on $d_j{}^t$ where j is the smallest integer for which

$$\xi_j f_j(x_1, \cdots, x_N) = \text{Max } (\xi_1 f_1, \cdots, \xi_k f_k) \tag{5.14}$$

For any ξ, let δ_ξ denote the Bayes solution given by the above rule. If, for every constant c, the set of samples $x_1, \cdots, x_N$ for which $f_i/f_j = c$ has the probability measure zero under every $F_l (i, j, l = 1, 2, \cdots, k)$, the Bayes solution δ_ξ is essentially unique; i.e., any other Bayes solution can differ from δ_ξ only on a set of ξ-measure zero. Suppose that the conditions of Theorem 5.4 are fulfilled and that for any ξ the decision function δ_ξ is (essentially) the only Bayes solution; then the problem of finding a minimax solution is reduced to the problem of finding a probability measure ξ^0 in Ω such that

$$r(F_1, \delta_{\xi^0}) = r(F_2, \delta_{\xi^0}) = \cdots = r(F_k, \delta_{\xi^0}) \tag{5.15}$$

A probability measure ξ^0 for which (5.15) holds must be a least favorable one, and δ_{ξ^0} is a minimax solution.[2]

As an illustration we shall discuss a few simple examples. Let $N = 2$, and let X_1 and X_2 be independently distributed with a common distribution. We shall assume that X_i can take only the values 0 and 1 and that Ω consists of two elements F_1 and F_2. Let the probability that $X_i = 1$ be equal to $\frac{1}{3}$ when F_1 is true, and equal to $\frac{2}{3}$ when F_2 is true. Furthermore, we put $W_{11} = W_{22} = 0$ and $W_{12} = W_{21} = 1$. We can verify that there exists a Bayes solution δ_{ξ^0} relative to the probability measure $\xi^0 = (\frac{1}{2}, \frac{1}{2})$, for which (5.15) holds. Hence ξ^0 is a least favorable a priori distribution. In order that a decision function δ be a Bayes solution relative to ξ^0 it is necessary and sufficient that

$$\delta_1(0, 0) = 1 \quad \text{and} \quad \delta_1(1, 1) = 0 \tag{5.16}$$

[2] See in this connection Theorem (17:D), page 161 of [55].

The values of $\delta_1(0, 1)$ and $\delta_1(1, 0)$ may be chosen arbitrarily. Clearly, not every Bayes solution relative to ξ^0 will be a minimax solution. For example, if we put $\delta_1(0, 1) = \delta_1(1, 0) = \alpha$, where α is a positive number $\neq \frac{1}{2}$, the resulting Bayes solution will not be a minimax solution. A minimax solution is given by the Bayes solution corresponding to $\delta_1(0, 1) = \delta_1(1, 0) = \frac{1}{2}$, as can easily be verified.

As a second example, consider the case where $N = 4$ and X_1, X_2, X_3, and X_4 are independently distributed with the same normal distribution having the variance $\sigma^2 = 4$. Suppose that Ω consists of three elements F_1, F_2, and F_3. The mean of the common normal distribution is -1 according to F_1, 0 according to F_2, and 1 according to F_3. The space D^t consists of three elements, $d_1{}^t$, $d_2{}^t$, and $d_3{}^t$; let $W_{ij} = 1$ for $i \neq j$, and $= 0$ for $i = j$. For any a priori distribution ξ, let δ_ξ denote the Bayes solution relative to ξ given by the following rule: $\delta_j(x_1, \cdots, x_4) = 1$, where j is the smallest integer for which $\xi_j{}^* = \text{Max } (\xi_1{}^*, \xi_2{}^*, \xi_3{}^*)$ and $\xi_i{}^*$ denotes the a posteriori probability that F_i is true after the sample $x_1, \cdots, x_4$ has been drawn. We can easily verify that $\xi_i{}^*(i = 1, 2, 3)$ depends only on the a priori distribution ξ and the arithmetic mean $\bar{x}$ of the observations. Furthermore we can easily verify that, if for some value of $\bar{x}$, say $\bar{x} = c$, we have $\xi_1{}^* = \text{Max } (\xi_1{}^*, \xi_2{}^*, \xi_3{}^*)$, then $\xi_1{}^* = \text{Max } (\xi_1{}^*, \xi_2{}^*, \xi_3{}^*)$ for any value $\bar{x} < c$. Similarly, if $\xi_3{}^* = \text{Max}_i\ (\xi_i{}^*)$ for $\bar{x} = c$, then $\xi_3{}^* = \text{Max}_i\ (\xi_i{}^*)$ for any $\bar{x} > c$. Hence there will be two constants c_1 and c_2 $(c_1 \leqq c_2)$ depending only on ξ such that $\xi_1{}^* = \text{Max}_i\ (\xi_i{}^*)$ if and only if $\bar{x} \leqq c_1$, and $\xi_3{}^* = \text{Max}_i\ (\xi_i{}^*)$ if and only if $\bar{x} \geqq c_2$.[3] Hence the decision function δ_ξ can be given as follows:

$$
\begin{aligned}
&\delta_1(x_1, \cdots, x_4) = 1 \qquad \text{when } \bar{x} \leqq c_1 \\
&\delta_2(x_1, \cdots, x_4) = 1 \qquad \text{when } c_1 < \bar{x} \leqq c_2 \\
&\delta_3(x_1, \cdots, x_4) = 1 \qquad \text{when } \bar{x} > c_2
\end{aligned}
\tag{5.17}
$$

Since the set of all samples for which $\bar{x}$ is equal to a given constant c is of measure zero, the Bayes solution δ_ξ is (essentially) unique.

Now let c_1 and c_2 be any two constants such that $c_1 \leqq c_2$, and let δ_{c_1,c_2} be the decision function given by (5.17). We can easily verify that there exists an a priori distribution ξ such that δ_ξ is identical to δ_{c_1,c_2}. To determine a ξ for which $\delta_\xi = \delta_{c_1,c_2}$, we have to solve the two equations in the components of ξ:

$$
\begin{aligned}
\xi_{1c_1}{}^* &= \text{Max } (\xi_{2c_1}{}^*, \xi_{3c_1}{}^*) \\
\xi_{3c_2}{}^* &= \text{Max } (\xi_{1c_2}{}^*, \xi_{2c_2}{}^*)
\end{aligned}
\tag{5.18}
$$

[3] The constants c_1 and c_2 may take the improper values $-\infty$ and ∞.

where ξ_{ic}^* denotes the a posteriori probability of F_i when $\bar{x} = c$.[4] Thus the class of all Bayes solutions coincides with the class of all decision functions δ_{c_1,c_2} corresponding to all possible values of c_1 and c_2 $(c_1 \leqq c_2)$.[5] Hence the class of all decision functions δ_{c_1,c_2} corresponding to all possible values of c_1 and c_2 is a complete class.

To find a minimax solution of our decision problem, we have merely to find two values c_1 and c_2 such that

$$c_1 \leqq c_2$$

$$r(F_1, \delta_{c_1,c_2}) = r(F_2, \delta_{c_1,c_2}) = r(F_3, \delta_{c_1,c_2}) \tag{5.19}$$

With the help of a table of the normal distribution the values c_1 and c_2 [satisfying (5.19)] can easily be found. They are equal to -0.803 and 0.803, respectively. Solving the equations (5.18) corresponding to these values of c_1 and c_2, we obtain the corresponding a priori distribution $\xi = (0.203, 0.593, 0.203)$, which is a least favorable a priori distribution for our problem.[6]

5.1.2 Non-Sequential Tests of a Hypothesis when Ω Is a Parametric Family of Distribution Functions

In this section we shall consider the case where the elements F of Ω can be described by a finite number of parameters, $\theta_1, \cdots, \theta_k$ (say). Then each element F of Ω can be represented by a point $\theta = (\theta_1, \cdots, \theta_k)$, called a parameter point, in the k-dimensional Cartesian space. The totality of all possible parameter points θ is called the parameter space. Since there is a one-to-one correspondence between the elements of Ω and the elements of the parameter space, it will cause no confusion if we use the symbol Ω to denote the parameter space also.

For the sake of simplicity, we shall restrict ourselves to problems when the elements F of Ω are absolutely continuous. As in Section 5.1.1, we shall consider only decision functions δ for which the probability is 1 that experimentation is carried out in a single stage by observing the values of the first N chance variables $X_1, \cdots, X_N$. Let $f(x_1, \cdots, x_N \mid \theta)$ denote the joint density function of $X_1, \cdots, X_N$ when θ is the true parameter point.

[4] We can easily verify that (5.18) always has a solution. If $c_1 < c_2$, the solution is unique.

[5] This characterization of the class of all Bayes solutions is contained as a special case in a more general characterization theorem given by Sobel [48].

[6] Since the Bayes solution relative to a given ξ is essentially unique, the above minimax solution must be admissible. The components of the least favorable a priori distribution, as given above, do not add up precisely to one due to rounding off errors.

Let ω be a given subset of the parameter space Ω, and suppose that the problem is to test the hypothesis H that the true parameter point θ is included in ω. Then the space D^t consists of two elements $d_1{}^t$ and $d_2{}^t$ (say), where $d_1{}^t$ denotes the decision to accept H and $d_2{}^t$ denotes the decision to reject H.

The weight function $W(\theta, d^t)$ is assumed to be a non-negative function such that $W(\theta, d_1{}^t) = 0$ for any θ in ω and $W(\theta, d_2{}^t) = 0$ for any θ in $\bar{\omega} = \Omega - \omega$. For the sake of simplicity, we shall consider here only simple weight functions $W(\theta, d^t)$, i.e., weight functions which can take only the values 0 and 1. Let ω_a denote the set of all points θ for which $W(\theta, d_2{}^t) = 1$. Clearly ω_a is a subset of ω. We shall refer to ω_a as the zone of preference for acceptance of H. Furthermore, let ω_r denote the set of all points θ for which $W(\theta, d_1{}^t) = 1$. Clearly ω_r is a subset of $\bar{\omega} = \Omega - \omega$, and we shall refer to it as the zone of preference for rejection of H. The set $\omega_I = \Omega - \omega_a - \omega_r$ is called the indifference zone. Clearly $W(\theta, d_1{}^t) = W(\theta, d_2{}^t) = 0$ for any θ in ω_I.

A decision function δ can be represented by a real-valued function $\delta(x_1, \cdots, x_N)$ such that $0 \leqq \delta(x_1, \cdots, x_N) \leqq 1$ for all values $x_1, \cdots, x_N$. The decision procedure is then given as follows: After the observations $x_1, \cdots, x_N$ have been made, perform a chance experiment with two possible outcomes, 1 and 2 (say), such that the probability of the outcome 1 is equal to $\delta(x_1, \cdots, x_N)$. Accept H if the outcome of this chance experiment is 1, and reject H if the outcome is 2.

Since the cost of experimentation is independent of δ, we can disregard it. The risk $r(\theta, \delta)$ is then given by

$$
\begin{aligned}
(5.20) \qquad r(\theta, \delta) &= \int_{M_N} f(x \mid \theta)\delta(x)\, dx && \text{if } \theta \text{ is in } \omega_r \\
&= \int_{M_N} f(x \mid \theta)[1 - \delta(x)]\, dx && \text{if } \theta \text{ is in } \omega_a \\
&= 0 && \text{if } \theta \text{ is in } \omega_I
\end{aligned}
$$

Here M_N denotes the totality of all samples $(x_1, \cdots, x_N)$, and x stands for $(x_1, \cdots, x_N)$. In the Neyman-Pearson theory, the risk $r(\theta, \delta)$ for θ in ω_a is called the size of the test, and $1 - r(\theta, \delta)$ for θ in ω_r is called the power of the test.

Any probability measure ξ on Ω can be given by a cumulative distribution function $\xi(\theta)$ in Ω. If $\xi(\theta)$ is the a priori distribution and if $x = (x_1, \cdots, x_N)$ is the observed sample, the a posteriori probability

of any subset ω of Ω is given by

$$P(\omega \mid \xi, x) = \frac{\int_{\omega} f(x \mid \theta)\, d\xi(\theta)}{\int_{\Omega} f(x \mid \theta)\, d\xi(\theta)} \tag{5.21}$$

A necessary and sufficient condition for a decision function δ to be a Bayes solution relative to a given a priori distribution ξ is that

$$\delta(x) = 1 \quad \text{whenever } P(\omega_a \mid \xi, x) > P(\omega_r \mid \xi, x) \tag{5.22}$$

and

$$\delta(x) = 0 \quad \text{whenever } P(\omega_a \mid \xi, x) < P(\omega_r \mid \xi, x) \tag{5.23}$$

except perhaps on a set of ξ-measure zero.[7] Since $P(\omega \mid \xi, x)$ is proportional to, $\int_{\omega} f(x \mid \theta)\, d\xi(\theta)$, conditions (5.22) and (5.23) are equivalent to

$$\delta(x) = 1 \quad \text{whenever} \int_{\omega_a} f(x \mid \theta)\, d\xi(\theta) > \int_{\omega_r} f(x \mid \theta)\, d\xi(\theta) \tag{5.24}$$

and

$$\delta(x) = 0 \quad \text{whenever} \int_{\omega_a} f(x \mid \theta)\, d\xi(\theta) < \int_{\omega_r} f(x \mid \theta)\, d\xi(\theta) \tag{5.25}$$

respectively. For any ξ, let δ_ξ denote the particular Bayes solution for which

$$\delta(x) = 1 \quad \text{when} \int_{\omega_a} f(x \mid \theta)\, d\xi(\theta) > \int_{\omega_r} f(x \mid \theta)\, d\xi(\theta) \tag{5.26}$$

and

$$\delta(x) = 0 \quad \text{for all other points } x \tag{5.27}$$

If the set of points x for which

$$\int_{\omega_a} f(x \mid \theta)\, d\xi(\theta) = \int_{\omega_r} f(x \mid \theta)\, d\xi(\theta) \tag{5.28}$$

holds is of ξ-measure zero, the Bayes solution is (essentially) unique; i.e., any other Bayes solution δ relative to ξ can differ from δ_ξ only on a set of ξ-measure zero.

In order to make the theory developed in Chapter 3 applicable to the problem studied in this section, we shall impose some restrictions on the class of density functions $f(x \mid \theta)$ which will guarantee the valid-

[7] By the ξ-measure of a subset R of the sample space we mean $\int_{\Omega}\int_{R} f(x \mid \theta)\, dx\, d\xi$.

ity of Assumption 3.7 postulated in Chapter 3. Assumptions 3.1 to 3.6 hold without any restrictions on $f(x, \theta)$. We shall formulate the following assumption.

Assumption 5.1. If $\{\theta^i\}$ $(i = 0, 1, 2, \cdots, ad\ inf.)$ *is a sequence of parameter points such that* $\lim_{i=\infty} \theta^i = \theta^0$, *then*

$$\lim_{i=\infty} \int_R f(x \mid \theta^i)\, dx = \int_R f(x \mid \theta^0)\, dx \tag{5.29}$$

uniformly in all subsets R *of the sample space* M_N.

In some problems it may not be easy to see whether Assumption 5.1 holds. The following lemma is useful in this connection.

Lemma 5.1. If $f(x \mid \theta)$ *is continuous in* θ, *Assumption 5.1 holds.*

This lemma is an immediate consequence of some results obtained by Robbins [43]. In fact, Robbins [43] has shown that, for any sequence $\{f_i(x)\}$ $(i = 0, 1, 2, \cdots,$ ad inf.) of density functions,

$$\lim_{i=\infty} f_i(x) = f_0(x) \tag{5.30}$$

in measure is equivalent to

$$\lim_{i=\infty} \int_R f_i(x)\, dx = \int_R f_0(x)\, dx \tag{5.31}$$

uniformly in all Borel sets R. We say that $\lim_{i=\infty} f_i(x) = f_0(x)$ in measure if, for every $\epsilon > 0$ and for every R with finite Borel measure, the set $S_i(R, \epsilon)$ of all x in R for which

$$| f_i(x) - f_0(x) | > \epsilon \tag{5.32}$$

satisfies the relation

$$\lim_{i=\infty} \int_{S_i(R,\epsilon)} dx = 0 \tag{5.33}$$

As Robbins [43] remarks, it can be shown that

$$\lim_{i=\infty} f_i(x) = f_0(x) \tag{5.34}$$

almost everywhere implies (5.30) but not conversely, and that

$$\lim_{i=\infty} \int_{M_N} | f_i(x) - f_0(x) |\, dx = 0 \tag{5.35}$$

is equivalent to (5.31).

Since $r(\theta, \delta) = 0$ identically in δ for any θ in ω_I, we can disregard the indifference zone ω_I and Ω can be replaced by the set-theoretical sum of ω_a and ω_r. Thus, in what follows, by the parameter space Ω we shall mean the set-theoretical sum of ω_a and ω_r. Furthermore we shall consider only cumulative distribution functions $\xi(\theta)$ for which

$$\int_{\omega_a} d\xi(\theta) + \int_{\omega_r} d\xi(\theta) = 1$$

We shall now formulate two more assumptions.

Assumption 5.2. *ω_a and ω_r are closed subsets of the k-dimensional Cartesian space.*

Assumption 5.3. *Ω is a bounded subset of the k-dimensional Cartesian space.*

We can easily verify that Assumptions 5.1, 5.2, and 5.3 imply the validity of Assumption 3.7 of Chapter 3.

The definition of regular convergence in the space of all decision functions δ, as given in Section 3.1.4, reduces in the case considered in this section to the following: Let $\{\delta_i\}$ $(i = 0, 1, 2, \cdots, \text{ad inf.})$ be a sequence of decision functions. We say that $\lim_{i=\infty} \delta_i = \delta_0$ (in the regular sense) if

$$\lim_{i=\infty} \int_R \delta_i(x)\, dx = \int_R \delta_0(x)\, dx \tag{5.36}$$

for any bounded subset R of the sample space M_N.

Let C_1 be the class of all decision functions δ for which (5.24) and (5.25) hold for some ξ (except perhaps on a set of x's with ξ-measure zero). Furthermore, let C_2 be the class of all decision functions δ for which (5.26) and (5.27) are fulfilled for some ξ. Clearly C_2 is a subclass of C_1. Let $\overline{C}_i$ be the closure of C_i $(i = 1, 2)$ in the sense of the convergence definition given in (5.36).

Theorem 3.20 of Chapter 3 yields the following two theorems.

Theorem 5.5. *If Assumptions 5.1 to 5.3 hold, C_1 is a complete class of decision functions.*

Theorem 5.6. *If Assumptions 5.1 to 5.3 hold, and if for any $\xi(\theta)$ the set of sample points x satisfying (5.28) is of Lebesgue measure zero, then C_2 is a complete class of decision functions.*

Theorem 3.19 of Chapter 3 yields the following theorems.

Theorem 5.7. The class $\overline{C}_1$ has the following property: For any δ not in $\overline{C}_1$ there exists an element δ^* of $\overline{C}_1$ such that $r(\theta, \delta^*) \leqq r(\theta, \delta)$ for all θ; i.e., $\overline{C}_1$ is essentially complete.

Theorem 5.8. If for any $\xi(\theta)$ the set of sample points x satisfying (5.28) is of Lebesgue measure zero, $\overline{C}_2$ has the following property: For any δ not in $\overline{C}_2$, there exists an element δ^* of $\overline{C}_2$ such that $r(\theta, \delta^*) \leqq r(\theta, \delta)$ for all θ; i.e., $\overline{C}_2$ is essentially complete.

We shall now discuss briefly a few simple examples. Let $X_1, X_2, \cdots, X_N$ be independently distributed with a common normal distribution having unit variance. Suppose that the mean θ is unknown and we wish to test the hypothesis that $\theta \leqq 0$. Let the set ω_a be given by the inequality $\theta \leqq -\rho$, and the set ω_r by the inequality $\theta \geqq \rho$, where ρ is a given positive number. In this case the necessary and sufficient conditions (5.24) and (5.25) for a decision function δ to be a Bayes solution relative to a given a priori distribution $\xi(\theta)$ reduce to

$$\delta(x_1, \cdots, x_N) = 1$$

whenever

$$\int_{-\infty}^{-\rho} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) > \int_{\rho}^{\infty} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) \tag{5.37}$$

and

$$\delta(x_1, \cdots, x_N) = 0$$

whenever

$$\int_{-\infty}^{-\rho} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) < \int_{\rho}^{\infty} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) \tag{5.38}$$

where $\bar{x}$ denotes the arithmetic mean of the observations $x_1, \cdots, x_N$. One can easily verify that there exists a constant c_0, depending only on ξ, such that the inequalities in (5.37) and (5.38) are equivalent to $\bar{x} < c_0$ and $\bar{x} > c_0$, respectively. The constant c_0 is equal to $-\infty$ when $\int_{-\infty}^{-\rho} d\xi(\theta) = 0$, and to ∞ when $\int_{\rho}^{\infty} d\xi(\theta) = 0$.

For any constant c, let $\delta_c(x_1, \cdots, x_N)$ denote the decision function given as follows: $\delta_c(x_1, \cdots, x_N) = 1$ when $\bar{x} < c$, and $\delta_c(x_1, \cdots, x_N) = 0$ when $\bar{x} \geqq c$. Since the set of sample points for which $\bar{x} = c$ is of Lebesgue measure zero, any Bayes solution must be (essentially) identical with $\delta_c(x_1, \cdots, x_N)$ for some value c. The converse is also true, as can easily be verified; i.e., for any given c, there exists an a priori distribution $\xi(\theta)$ such that $\delta_c(x_1, \cdots, x_N)$ is a Bayes solution relative to $\xi(\theta)$.

We can even find an a priori distribution $\xi(\theta)$ with the above property and such that the set consisting of the two points $\theta = \pm\rho$ has probability 1. Thus the class C_2 in Theorem 5.6 coincides with the class of all decision functions $\delta_c(x_1, \cdots, x_N)$ corresponding to all possible values of c.[8] Since any decision function δ which is a limit of a sequence of members of C_2 is itself a member of C_2, we have $\bar{C}_2 = C_2$. Applying Theorem 5.8, we arrive at the following result: *For any decision function δ there exists a constant c such that*

$$r(\theta, \delta_c) \leqq r(\theta, \delta) \tag{5.39}$$

for all θ.

Let $\xi_0(\theta)$ be the cumulative distribution function that assigns the probability $\frac{1}{2}$ to each of the points $\theta = \pm\rho$. Furthermore let δ_0 be the decision function for which $\delta_0 = 1$ if $\bar{x} < 0$ and $\delta_0 = 0$ if $\bar{x} \geqq 0$. Then δ_0 is a Bayes solution relative to ξ_0. Since

$$r(\rho, \delta_0) = r(-\rho, \delta_0) = \operatorname{Max}_\theta r(\theta, \delta_0) \tag{5.40}$$

$\xi_0(\theta)$ is a least favorable a priori distribution and δ_0 is a minimax solution.

As a second example, consider the case where $X_1, \cdots, X_N$ are again independently distributed with a common normal distribution having variance 1, but the hypothesis H to be tested is that the unknown mean θ lies in the interval $(-\rho, \rho)$, where ρ is a positive number. Let the zone ω_a be given by the inequality $|\theta| \leqq \rho_1$, and the zone ω_r by the inequality $|\theta| \geqq \rho_2$, where ρ_1 and ρ_2 are given positive numbers such that $\rho_1 < \rho < \rho_2$. A necessary and sufficient condition for a decision function δ to be a Bayes solution relative to a given a priori distribution $\xi(\theta)$ will now be given by

$$\delta(x_1, \cdots, x_N) = 1$$

whenever

$$\int_{|\theta| \geqq \rho_2} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) < \int_{|\theta| \leqq \rho_1} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) \tag{5.41}$$

and

$$\delta(x_1, \cdots, x_N) = 0$$

whenever

$$\int_{|\theta| \geqq \rho_2} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) > \int_{|\theta| \leqq \rho_1} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) \tag{5.42}$$

[8] This characterization of the class of all Bayes solutions for the given problem is contained as a special case in a characterization theorem by Sobel relating to a more general class of problems [48].

except perhaps on a set of Lebesgue measure zero. Let $\psi_1(\bar{x})$ denote the integral on the left-hand side of the inequality in (5.41), and $\psi_2(\bar{x})$ the integral on the right-hand side of the same inequality. We can easily verify that both $\psi_1(\bar{x})$ and $\psi_2(\bar{x})$ can be differentiated under the integral sign with respect to $\bar{x}$ any number of times. Thus

$$\begin{aligned}\frac{d^2\psi_1(\bar{x})}{d\bar{x}^2} &= N^2 \int_{|\theta| \geqq \rho_2} \theta^2 e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) \\ \frac{d^2\psi_2(\bar{x})}{d\bar{x}^2} &= N^2 \int_{|\theta| \leqq \rho_1} \theta^2 e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta)\end{aligned} \tag{5.43}$$

It follows immediately from (5.43) that

$$\frac{d^2\psi_1(\bar{x})}{d\bar{x}^2} > \frac{d^2\psi_2(\bar{x})}{d\bar{x}^2} \tag{5.44}$$

whenever $\psi_1(\bar{x}) \geqq \psi_2(\bar{x})$. Clearly, if $\int_{|\theta| \geqq \rho_2} d\xi(\theta) > 0$ and $\int_{|\theta| \leqq \rho_1} d\xi(\theta) > 0$,

$$\begin{aligned}\lim_{\bar{x} = \infty} [\psi_1(\bar{x}) - \psi_2(\bar{x})] &= \infty \\ \lim_{\bar{x} = -\infty} [\psi_1(\bar{x}) - \psi_2(\bar{x})] &= \infty\end{aligned} \tag{5.45}$$

It follows from (5.44) and (5.45) that there exist two constants c_1 and c_2 such that $c_1 \leqq c_2$ and

$$\psi_1(\bar{x}) - \psi_2(\bar{x}) > 0 \tag{5.46}$$

when $\bar{x} < c_1$ and $\bar{x} > c_2$, and

$$\psi_1(\bar{x}) - \psi_2(\bar{x}) < 0 \tag{5.47}$$

when $c_1 < \bar{x} < c_2$. The constants c_1 and c_2 may take the improper values $-\infty$ and $+\infty$. For example, if $\int_{|\theta| \leqq \rho_1} d\xi(\theta) = 1$, then $c_1 = -\infty$, $c_2 = \infty$, and $\psi_1(\bar{x}) - \psi_2(\bar{x}) < 0$ for all $\bar{x}$. Similarly, if $\int_{|\theta| \geqq \rho_2} d\xi(\theta) = 1$, then c_1 and c_2 are equal to ∞ and $\psi_1(\bar{x}) - \psi_2(\bar{x}) > 0$ for all $\bar{x}$.

For any constants c_1 and c_2 ($c_1 \leqq c_2$) let $\delta_{c_1,c_2}(x_1, \cdots, x_N)$ denote the decision function given as follows: $\delta_{c_1,c_2}(x_1, \cdots, x_N) = 0$ when $\bar{x} \leqq c_1$ or $\bar{x} \geqq c_2$; $\delta_{c_1,c_2}(x_1, \cdots, x_N) = 1$ for all other samples. Since the set of sample points $(x_1, \cdots, x_N)$ for which $\bar{x}$ is equal to a constant c is of Lebesgue measure zero, the Bayes solution relative to a given a

priori distribution $\xi(\theta)$ is (essentially) unique and coincides with δ_{c_1,c_2} for some values of c_1 and c_2. Conversely, as can easily be shown, for any c_1 and c_2 there exists an a priori distribution $\xi(\theta)$ such that δ_{c_1,c_2} is a Bayes solution relative to $\xi(\theta)$. Such an a priori distribution exists even if we restrict ourselves to distributions $\xi(\theta)$ which assign the probability 1 to the set consisting of the three points $\theta = -\rho_2, 0, \rho_2$. Hence the class C_2 in Theorem 5.6 is identical to the class of all decision functions δ_{c_1,c_2} corresponding to all possible values of c_1 and c_2.[9] The closure $\overline{C}_2$ of C_2 is obviously equal to C_2. Hence Theorem 5.8 yields the result: *For any decision function δ there exist two constants c_1 and c_2 such that $r(\theta, \delta_{c_1,c_2}) \leqq r(\theta, \delta)$ for all θ.*[10]

For any positive value $\alpha < \frac{1}{2}$, let $\xi_\alpha(\theta)$ denote the a priori distribution which assigns the probability α to each of the points $\theta = -\rho_1$ and $\theta = \rho_1$, and the probability $\frac{1}{2} - \alpha$ to each of the points $\theta = -\rho_2$ and $\theta = \rho_2$. It follows from reasons of symmetry that there exists a constant c_α, depending only on α, such that a Bayes solution relative to ξ_α is given by $\delta_{-c_\alpha,c_\alpha}$. Furthermore it is clear that

$$r(-\theta, \delta_{-c_\alpha,c_\alpha}) = r(\theta, \delta_{-c_\alpha,c_\alpha}) \tag{5.48}$$

and

$$\text{Max}_\theta\, r(\theta, \delta_{-c_\alpha,c_\alpha}) = \text{Max}\,[r(\rho_1, \delta_{-c_\alpha,c_\alpha}), r(\rho_2, \delta_{-c_\alpha,c_\alpha})] \tag{5.49}$$

We can easily verify that there exists a value α_0 such that

$$r(\rho_1, \delta_{-c_{\alpha_0},c_{\alpha_0}}) = r(\rho_2, \delta_{-c_{\alpha_0},c_{\alpha_0}}) \tag{5.50}$$

Clearly ξ_{α_0} is a least favorable a priori distribution, and $\delta_{-c_{\alpha_0},c_{\alpha_0}}$ is a minimax solution.

5.1.3 Non-Sequential Point and Interval Estimation when Ω Is a Parametric Family of Distribution Functions

As in Section 5.1.2, we shall again assume that any element F of Ω can be represented by a parameter point $\theta = (\theta_1, \cdots, \theta_k)$ in the k-dimensional Cartesian space, and we shall use the symbol Ω to denote the parameter space. Again we shall consider only decision functions δ according to which experimentation is carried out in a single stage by observing the values of $X_1, \cdots, X_N$. For the sake of simplicity we shall assume that the elements F of Ω are absolutely continuous and

[9] This characterization of the class of all Bayes solutions, together with other more general results, is contained in a paper by Sobel [48].

[10] A similar result was obtained by Lehmann [30] in the case of testing the hypothesis that θ is equal to a specified value θ_0.

let $f(x_1, \cdots, x_N \mid \theta)$ denote the joint density function of $X_1, \cdots, X_N$ when θ is the true parameter point.

We shall consider first the problem of point estimation. For any parameter point θ^*, let $d_{\theta^*}{}^t$ denote the terminal decision to estimate the true parameter point θ by θ^*. The space D^t consists of all elements $d_{\theta^*}{}^t$ corresponding to all possible parameter points θ^*. We shall use the symbol $W(\theta, \theta^*)$ to denote $W(\theta, d_{\theta^*}{}^t)$; i.e., $W(\theta, \theta^*)$ is the loss suffered if θ is the true parameter point and our point estimate of θ is θ^*.

A decision function δ can now be represented by a function $\delta(\theta \mid x_1, \cdots, x_N)$ of θ, $x_1, \cdots, x_N$ such that for any given sample $x_1, \cdots, x_N$ the function $\delta(\theta \mid x_1, \cdots, x_N)$ is a cumulative distribution function in Ω. After the sample $(x_1, \cdots, x_N)$ has been obtained, the actual selection of a point estimate θ^* is made with the help of a chance mechanism constructed in such a way that the probability that θ^* is included in a subset ω of Ω is equal to $\int_\omega d\delta(\theta \mid x_1, \cdots, x_N)$. Since the cost of experimentation is independent of δ and can, therefore, be disregarded, the risk is given by

$$r(\theta, \delta) = \int_{M_N} \left[\int_\Omega W(\theta, \theta^*)\, d\delta(\theta^* \mid x) \right] f(x \mid \theta)\, dx \tag{5.51}$$

where $x = (x_1, \cdots, x_N)$ and M_N denotes the sample space.

In Chapter 3 the assumption was made that D^t is compact. In order to guarantee the compactness of D^t, we shall make the following assumption.

Assumption 5.4. Ω is a bounded and closed subset of the k-dimensional Cartesian space, and $W(\theta, \theta^*)$ is continuous jointly in θ and θ^*.

This assumption, together with Assumption 5.1 formulated earlier, implies the validity of Assumptions 3.1 to 3.7 of Chapter 3.

For any probability distribution $\xi = \xi(\theta)$ in Ω, let

$$W(\xi, \theta^*) = \int_\Omega W(\theta, \theta^*)\, d\xi(\theta) \tag{5.52}$$

For any a priori distribution ξ and for any sample point x, let $\omega_{\xi,x}$ denote the totality of all parameter points θ for which

$$W(\xi_x, \theta) = \mathrm{Min}_{\theta^*} W(\xi_x, \theta^*) \tag{5.53}$$

where ξ_x denotes the a posteriori distribution in Ω after the sample x has been observed.

A necessary and sufficient condition for a decision function δ to be

a Bayes solution relative to the a priori distribution ξ is that

$$\int_{\omega_{\xi,x}} d\delta(\theta \mid x) = 1 \tag{5.54}$$

except perhaps on a set of sample points x whose ξ measure is zero.

The special case where, for any ξ and x, $\omega_{\xi,x}$ consists of a single element $\theta_{\xi,x}^*$ is of considerable interest. In this case, the Bayes solution relative to any given ξ is unique (except perhaps on a set of x's with ξ-measure zero) and is identical to the decision function $\delta_\xi(\theta \mid x_1, \cdots, x_N)$, which assigns the probability 1 to the parameter point $\theta_{\xi,x}^*$.

Since Assumptions 5.1 and 5.4 imply the validity of Assumptions 3.1 to 3.7, the results obtained in Chapter 3 yield the following theorem.

Theorem 5.9. If Assumptions 5.1 and 5.4 hold, then

(i) *A least favorable distribution $\xi(\theta)$ exists.*

(ii) *A minimax solution exists.*

(iii) *The class of all decision functions $\delta(\theta \mid x_1, \cdots, x_N)$ which satisfy (5.54) for some ξ (identically in x) is a complete class of decision functions.*

As an illustration, we shall consider the following example. Let $X_1, \cdots, X_N$ be independently distributed with the same normal distribution having variance = 1. The problem is to give a point estimate for θ. Let $W(\theta, \theta^*) = (\theta - \theta^*)^2$, and let the domain of θ be restricted to the finite interval $[a, b]$ $(a < b)$. In this example, for any ξ and any x the set $\omega_{\xi,x}$ will consist of a single point $\theta_{\xi,x}^*$. We have

$$\theta_{\xi,x}^* = \frac{\int_a^b \theta e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta)}{\int_a^b e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta)} = \psi_\xi(N\bar{x}) \quad \text{(say)} \tag{5.55}$$

Clearly

$$a \leqq \psi_\xi(N\bar{x}) \leqq b \tag{5.56}$$

for all $\bar{x}$.

We shall now derive a necessary and sufficient condition for a function $\psi(t)$ to be such that there exists a ξ for which $\psi(t) = \psi_\xi(t)$ for all real t. For any $\xi(\theta)$, let $\xi^*(\theta)$ be the cumulative distribution function for which

$$\int_\omega d\xi^*(\theta) = \frac{\int_\omega e^{-\frac{1}{2}N\theta^2}\, d\xi(\theta)}{\int_a^b e^{-\frac{1}{2}N\theta^2}\, d\xi(\theta)} \tag{5.57}$$

for any subset ω of Ω. Clearly, if $\psi(t) = \psi_\xi(t)$,

$$\psi(t) = \frac{\int_a^b \theta e^{t\theta}\, d\xi^*(\theta)}{\int_a^b e^{t\theta}\, d\xi^*(\theta)} = \frac{\phi'(t)}{\phi(t)} \tag{5.58}$$

where $\phi(t)$ is the moment-generating function of θ when $\xi^*(\theta)$ is the distribution of θ; i.e.,

$$\phi(t) = \int_a^b e^{t\theta}\, d\xi^*(\theta) \tag{5.59}$$

and $\phi'(t)$ is the derivative of $\phi(t)$ with respect to t.

Thus a necessary condition for $\psi(t)$ to be such that there exists a ξ for which $\psi(t) = \psi_\xi(t)$ is that $\psi(t) = \phi'(t)/\phi(t)$, where $\phi(t)$ is the moment-generating function of a chance variable with the range $[a, b]$. To show the sufficiency of this condition let us assume that

$$\psi(t) = \frac{\phi'(t)}{\phi(t)}$$

where

$$\phi(t) = \int_a^b e^{t\theta}\, d\xi_1(\theta)$$

and $\xi_1(\theta)$ is a cumulative distribution function in Ω. Then

$$\psi(t) = \frac{\int_a^b \theta e^{t\theta}\, d\xi_1(\theta)}{\int_a^b e^{t\theta}\, d\xi_1(\theta)} \tag{5.60}$$

Let $\xi_2(\theta)$ be the distribution function which satisfies the equation

$$\int_\omega d\xi_2(\theta) = \frac{\int_\omega e^{\frac{1}{2}N\theta^2}\, d\xi_1(\theta)}{\int_a^b e^{\frac{1}{2}N\theta^2}\, d\xi_1(\theta)} \tag{5.61}$$

It then follows from (5.60) that

$$\psi(t) = \frac{\int_a^b \theta e^{t\theta - \frac{1}{2}N\theta^2}\, d\xi_2(\theta)}{\int_a^b e^{t\theta - \frac{1}{2}N\theta^2}\, d\xi_2(\theta)} \tag{5.62}$$

Hence $\psi(t) = \psi_{\xi_2}(t)$, and the sufficiency of our condition is proved. Thus we arrive at the following necessary and sufficient conditions for a decision function $\delta(\theta \mid x_1, \cdots, x_N)$ to be such that there exists an a priori distribution ξ relative to which $\delta(\theta \mid x_1, \cdots, x_N)$ is a Bayes solution:

(5.63) $\delta(\theta \mid x_1, \cdots, x_N)$ assigns the probability 1 to $\theta = \psi(N\bar{x})$

(except perhaps on a set of Lebesgue measure zero), and

$$\psi(t) = \frac{\phi'(t)}{\phi(t)} \tag{5.64}$$

where $\phi(t)$ is the moment-generating function of some chance variable u with the range $[a, b]$.

Since Assumptions 5.1 and 5.4 are obviously fulfilled for our problem, it follows from Theorem 5.9 that the class of all decision functions δ which satisfy (5.63) and (5.64) is a complete class.

It is interesting to note that $\psi(t) = t/N$ does not satisfy (5.64), since $\phi'(t)/\phi(t) = t/N$ is possible only when $\phi(t)$ is the moment-generating function of a normally distributed chance variable, but not of a chance variable with finite range. Thus the decision function $\delta(\theta \mid x_1, \cdots, x_N)$ which assigns the probability 1 to $\bar{x}$ is not an admissible decision function. We can also show that $\delta(\theta \mid x_1, \cdots, x_N)$ is not an admissible decision function if it assigns the value 1 to $\bar{x}$ when $a \leqq \bar{x} < b$, to a when $\bar{x} < a$, and to b when $\bar{x} > b$. The reason for this somewhat surprising result is that we limited the range of θ to a finite interval. If no restrictions are imposed on the domain of θ, the decision function $\delta(\theta \mid x_1, \cdots, x_N)$ that assigns the probability 1 to $\theta = \bar{x}$ can be shown to be a minimax solution of the problem.

The following two interesting examples of minimax solutions are due to J. Hodges and E. L. Lehmann:

I. *Minimax Point Estimate of the Mean of a Binomial Variate.*[11] Let X be a binomial random variable, i.e.,

$$\text{prob. } \{X = x\} = \binom{N}{x} \theta^x (1 - \theta)^{N-x} \quad (x = 0, 1, \cdots, N)$$

Let $W(\theta, \theta^*) = (\theta - \theta^*)^2$. Then the minimax estimate of θ is

$$\theta^*(x) = \frac{1}{1 + \sqrt{N}} \left(\frac{1}{\sqrt{N}} x + \frac{1}{2} \right)$$

[11] H. Rubin found the minimax estimate of the mean of a binomial variate before Hodges and Lehmann did.

where x is the observed value of X. In other words, the minimax decision function $\delta(\theta \mid x)$ assigns the probability 1 to the parameter value $\theta^*(x)$. This can be shown as follows: It is easily checked that when

$$c_1 = \frac{1}{\sqrt{N}\,(1+\sqrt{N})} \qquad \text{and} \qquad c_2 = \frac{1}{2(1+\sqrt{N})}$$

then

$$E(c_1X + c_2 - \theta)^2 = \frac{1}{4(1+\sqrt{N})^2}$$

independent of θ. It is, therefore, sufficient to prove that the above estimate is a Bayes estimate. Straightforward calculation shows that the Bayes estimate corresponding to $d\xi(\theta) = C\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta$ $(\alpha, \beta > 0)$ is

$$\theta(x) = \frac{x+\alpha}{\alpha+\beta+N}$$

Hence, $\theta^*(x) = c_1x + c_2$ is the Bayes estimate corresponding to $\alpha = \beta = \sqrt{N}/2$.

II. *Minimax Point Estimate of the Mean of a Chance Variable with Range* [0, 1]. Let $X_1, \cdots, X_N$ be independently and identically distributed over the interval [0, 1]. The common distribution is assumed to be entirely unknown. Let $E(X_i) = \theta$, and suppose that the problem is to construct a point estimate of θ. Let $W(\theta, \theta^*) = (\theta - \theta^*)^2$. Then the minimax estimate of θ is

$$\theta^*(x) = \frac{1}{1+\sqrt{N}}\,(\sqrt{N}\bar{x} + \tfrac{1}{2})$$

where

$$\bar{x} = \frac{x_1 + \cdots + x_N}{N}$$

and x_i is the observed value of X_i. This can be seen as follows: Since the above estimate is the minimax estimate when the X's can take on only the values 0 and 1, it is sufficient to show that the risk of this estimate is bounded above by $1/[4(1+\sqrt{N})^2]$. This is easily verified, since

$$E(Nc_1\overline{X} + c_2 - \theta)^2 = N^2{c_1}^2{\sigma_{\overline{X}}}^2 + [(Nc_1 - 1)\theta + c_2]^2$$

But

$${\sigma_{\overline{X}}}^2 = \frac{1}{N}\,[E(X^2) - \theta^2] \leqq \frac{1}{N}\,(\theta - \theta^2)$$

which is the variance in the binomial case. Hence the risk takes on its maximum value in the binomial case.

We shall now discuss briefly the problem of interval estimation. For simplicity, we shall consider the case of a single unknown parameter θ; i.e., $k = 1$. For any closed interval I of the real axis, let $d_I{}^t$ denote the terminal decision to state that the true parameter value θ is included in I. The space D^t consists of all elements $d_I{}^t$ corresponding to all possible intervals I.[12] Any interval I is characterized by its midpoint θ^* and its length l. Let $W(\theta, \theta^*, l)$ be the loss suffered when θ is the true value and estimate θ to be included in the interval with midpoint θ^* and length l.

A decision function δ can be represented by a function $\delta(\theta, l \mid x_1, \cdots, x_N)$ such that for any given sample $x_1, \cdots, x_N$ the function $\delta(\theta, l \mid x_1, \cdots, x_N)$ is a cumulative distribution function in the space of all pairs (θ, l). After the sample $x_1, \cdots, x_N$ has been drawn, the actual selection of θ^* and l is made with the help of a chance mechanism such that for any given real values θ_0 and l_0 the probability that $\theta^* < \theta_0$ and $l < l_0$ is equal to $\delta(\theta_0, l_0 \mid x_1, \cdots, x_N)$.

Since the cost of experimentation may be disregarded, the risk is given by

$$r(\theta, \delta) = \int_{M_N} \left[\int_{\Omega^*} W(\theta, \theta^*, l)\, d\delta(\theta^*, l \mid x) \right] f(x \mid \theta)\, dx \tag{5.65}$$

where Ω^* denotes the space of all pairs (θ^*, l).

We shall make the following assumption.

Assumption 5.5. Ω is a bounded and closed subset of the real axis and $W(\theta, \theta^*, l)$ is continuous jointly in θ, θ^*, and l.

This assumption, together with Assumption 5.1, implies the validity of Assumptions 3.1 to 3.7 formulated in Chapter 3.

Let

$$W(\xi, \theta^*, l) = \int_{\Omega} W(\theta, \theta^*, l)\, d\xi(\theta) \tag{5.66}$$

For any a priori distribution ξ and for any sample $x = (x_1, \cdots, x_N)$, let $D_{\xi,x}$ denote the totality of all pairs (θ^*, l) for which

$$W(\xi_x, \theta^*, l) = \operatorname{Min}_{\bar{\theta}, \bar{l}} W(\xi_x, \bar{\theta}, \bar{l}) \tag{5.67}$$

where ξ_x denotes the a posteriori distribution in Ω after the sample x has been observed.

[12] In some problems the admissible intervals may be restricted to a certain subclass of the class of all intervals, such as the class of all intervals whose length does not exceed a given value, or whose length is equal to a given value, etc.

A decision function $\delta(\theta^*, l \mid x)$ is a Bayes solution relative to a given a priori distribution ξ if and only if

$$\int_{D_{\xi,x}} d\delta(\theta^*, l \mid x) = 1 \tag{5.68}$$

for all x, except perhaps on a set of ξ-measure zero. In the special case where for any x the set $D_{\xi,x}$ consists of a single element, the Bayes solution relative to ξ is unique (except perhaps on a set of ξ-measure zero).

Since Assumptions 3.1 to 3.7 of Chapter 3 follow from Assumptions 5.1 and 5.5, the results of Chapter 3 yield the following theorem.

Theorem 5.10. If Assumptions 5.1 and 5.5 hold, then
(i) *A least favorable distribution $\xi(\theta)$ exists.*
(ii) *A minimax solution exists.*
(iii) *The class of all decision functions $\delta(\theta, l \mid x_1, \cdots, x_N)$ which satisfy (5.68) for at least one ξ is a complete class of decision functions.*

Let us consider the following example: The chance variables $X_1, \cdots, X_N$ are independently distributed and they have the following common density function: $f(x_i \mid \theta) = 1$ for $-\frac{1}{2} + \theta \leqq x_i \leqq \frac{1}{2} + \theta$ and $= 0$ elsewhere. The mean θ is unknown, but the domain of θ is restricted to the closed interval $[a, b]$. For any probability distribution ξ in Ω and for any real numbers a' and b' $(a' \leqq b')$ for which the common part of $[a, b]$ and $[a', b']$ has a positive ξ-measure, let $\xi_{a',b'}$ denote the conditional distribution of θ when θ is restricted to the interval $[a', b']$. We then have, for any subset ω of the interval $[a', b']$,

$$\int_{\omega} d\xi_{a',b'}(\theta) = \frac{\int_{\omega} d\xi(\theta)}{\int_{a'}^{b'} d\xi(\theta)} \tag{5.69}$$

Let

$$u = \text{Min}\,(x_1, \cdots, x_N) \quad \text{and} \quad v = \text{Max}\,(x_1, \cdots, x_N) \tag{5.70}$$

If ξ is the a priori distribution in Ω, the a posteriori distribution after the sample x has been drawn is given by

$$\xi_x = \xi_{v-\frac{1}{2},u+\frac{1}{2}} \tag{5.71}$$

as can easily be verified.

To simplify our problem, we shall admit only a single value l_0 for $l (0 < l_0 < \frac{1}{2})$ and restrict the choice of the experimenter to the

choice of the midpoint θ^*. The weight function $W(\theta, \theta^*, l_0)$ is assumed to be given as follows:

$$
\begin{aligned}
(5.72)\quad W(\theta, \theta^*, l_0) &= 0 && \text{if } \theta^* - l_0 \leqq \theta \leqq \theta^* + l_0 \\
&= (\theta - \theta^* + l_0)^2 && \text{if } \theta < \theta^* - l_0 \\
&= (\theta - \theta^* - l_0)^2 && \text{if } \theta > \theta^* + l_0
\end{aligned}
$$

For any distribution $\xi(\theta)$ in Ω we then have

$$
(5.73)\quad W(\xi, \theta^*, l_0) = \int_{\theta < \theta^* - l_0} (\theta - \theta^* + l_0)^2 \, d\xi(\theta) + \int_{\theta > \theta^* + l_0} (\theta - \theta^* - l_0)^2 \, d\xi(\theta)
$$

For any ξ, u, v, let $\omega_{\xi,u,v}$ denote the totality of all values θ^* for which $W(\xi_{v-\frac{1}{2},u+\frac{1}{2}}, \theta^*, l_0)$ becomes a minimum. Then the necessary and sufficient condition for a Bayes solution given in (5.68) reduces to [13]

$$
(5.74)\quad \int_{\omega_{\xi,u,v}} d\delta(\theta \mid x_1, \cdots, x_N) = 1
$$

for all $x = (x_1, \cdots, x_N)$ except perhaps on a set of ξ-measure zero.

There does not seem to be a simple characterization of the class of all the Bayes solutions. Instead we shall study the Bayes solutions relative to the uniform a priori distribution which have some interesting properties, as will be seen below.

Let $\xi^0(\theta)$ be the uniform a priori distribution; i.e.,

$$
(5.75)\quad \xi^0(\theta) = \frac{\theta - a}{b - a}
$$

for any θ in the interval $[a, b]$. Let $\theta_{u,v}$ be the midpoint of the intersection of the intervals $[v - \frac{1}{2}, u + \frac{1}{2}]$ and $[a, b]$. If the length of this intersection is greater than or equal to $2l_0$, $\omega_{\xi^0,u,v}$ consists of the single point $\theta_{u,v}$, as can easily be verified. If the length of the intersection is less than $2l_0$, $\omega_{\xi^0,u,v}$ consists of all points θ^* for which the interval $[\theta^* - l_0, \theta^* + l_0]$ covers the intersection in question. For any θ in the interval $[a + 1, b - 1]$, the probability is 1 that $[v - \frac{1}{2}, u + \frac{1}{2}]$ is contained in $[a, b]$. Then, if the length of the interval $[v - \frac{1}{2}, u + \frac{1}{2}]$ is greater than or equal to $2l_0$, $\theta_{u,v}$ is equal to $(u + v)/2$.

Let $\delta_0(\theta \mid x_1, \cdots, x_N)$ be a Bayes solution relative to ξ^0. It can be shown that $r(\theta, \delta_0)$ is constant over the θ-interval $[a + 1, b - 1]$. Furthermore we can easily verify that $\mathrm{Max}_\theta \, r(\theta, \delta_0)$ is equal to the

[13] The argument l in $\delta(\theta, l \mid x_1, \cdots, x_N)$ is dropped, since l is restricted to a fixed value l_0.

constant value of $r(\theta, \delta_0)$ in the interval $[a + 1, b - 1]$. From this it follows that

$$\lim_{(b-a) \to \infty} [\mathrm{Max}_\theta\, r(\theta, \delta_0) - r(\xi^0, \delta_0)] = 0 \tag{5.76}$$

Hence

$$\lim_{(b-a) \to \infty} [\mathrm{Max}_\theta\, r(\theta, \delta_0) - \mathrm{Inf}_\delta\, \mathrm{Max}_\theta\, r(\theta, \delta)] = 0 \tag{5.77}$$

and

$$\lim_{(b-a) \to \infty} [\mathrm{Inf}_\delta\, r(\xi^0, \delta) - \mathrm{Sup}_\xi\, \mathrm{Inf}_\delta\, r(\xi, \delta)] = 0 \tag{5.78}$$

Thus for sufficiently large $b - a$ the distribution ξ^0 is for all practical purposes a least favorable distribution and δ_0 is for all practical purposes a minimax solution.

5.1.4. Non-Sequential Decision Problems when D^t Is Finite and Ω Is a Parametric Class of Distribution Functions

The case where D^t consists of two elements was treated in Section 5.1.2. Here we shall deal with the case where D^t is finite and contains more than two elements. Let $d_1{}^t, \cdots, d_u{}^t$ be the elements of D^t $(u > 2)$. As before, we shall admit only decision functions δ according to which experimentation is carried out in a single stage by observing the values of $X_1, \cdots, X_N$. For any parameter point $\theta = (\theta_1, \cdots, \theta_k)$ we shall assume that the corresponding joint distribution of $X_1, \cdots, X_N$ is absolutely continuous. Let $f(x_1, \cdots, x_N \mid \theta)$ denote the joint density function of $X_1, \cdots, X_N$ when θ is the true parameter point. Let

$$W_i(\theta) = W(\theta, d_i{}^t) \tag{5.79}$$

i.e., $W_i(\theta)$ is the loss when θ is the true parameter point and the decision $d_i{}^t$ is made.

Any decision function δ can be represented by a vector function $\delta(x_1, \cdots, x_N)$ with the components $\delta_1(x_1, \cdots, x_N), \cdots, \delta_u(x_1, \cdots, x_N)$ satisfying the relations

$$\delta_i(x_1, \cdots, x_N) \geqq 0 \qquad \text{and} \qquad \sum_{i=1}^{u} \delta_i(x_1, \cdots, x_N) = 1 \tag{5.80}$$

Here $\delta_i(x_1, \cdots, x_N)$ is equal to the probability that we shall decide on $d_i{}^t$ when $(x_1, \cdots, x_N)$ is the observed sample. The cost of experimentation being disregarded, the risk when θ is true and δ is adopted is given by

$$r(\theta, \delta) = \sum_{i=1}^{u} \int_{M_N} W_i(\theta)\, \delta_i(x) f(x \mid \theta)\, dx \tag{5.81}$$

where $x = (x_1, \cdots, x_N)$ and M_N denotes the sample space (totality of all samples x).

For any probability distribution $\xi(\theta)$ in Ω, let $W_i(\xi)$ denote the expected value of $W_i(\theta)$; i.e.,

$$W_i(\xi) = \int_\Omega W_i(\theta)\, d\xi(\theta) \tag{5.82}$$

For any a priori distribution ξ, let ξ_x denote the a posteriori distribution in Ω after the sample x has been observed. We shall refer to the quantity $W_i(\xi_x)$ as the a posteriori risk associated with the decision d_i^t.

A necessary and sufficient condition for a decision function δ to be a Bayes solution relative to a given a priori distribution ξ is given as follows: For all sample points x (except perhaps on a set of ξ-measure zero) we have

$$\delta_i(x) = 0 \tag{5.83}$$

whenever $W_i(\xi_x) > W_j(\xi_x)$ for some j.

The proof of the above statement is very simple and is omitted. We shall now formulate the following assumption.

Assumption 5.6. Ω is a closed and bounded subset of the k-dimensional Cartesian space and $W_i(\theta)$ $(i = 1, \cdots, u)$ is a continuous function of θ.

Assumption 5.6, together with Assumption 5.1 formulated earlier (Section 5.1.2), implies the validity of Assumptions 3.1 to 3.7. Assumptions 3.1 to 3.6 hold when $W_i(\theta)$ is a bounded function of θ $(i = 1, 2, \cdots, u)$ without any restriction on Ω and $f(x \mid \theta)$.

Let C denote the class of all Bayes solutions; i.e., a decision function δ is an element of C if and only if there exists an a priori distribution ξ such that δ is a Bayes solution relative to ξ. Also let $\bar{C}$ denote the closure of C [in the sense of the convergence definition given in (5.36)]. Applying the results of Chapter 3, we obtain the following two theorems:

Theorem 5.11. If Assumptions 5.1 and 5.6 hold, then

(i) *There exists a least favorable a priori distribution.*

(ii) *There exists a minimax solution which must be a Bayes solution relative to any least favorable a priori distribution.*

(iii) *C is a complete class of decision functions.*

Theorem 5.12. If $W_i(\theta)$ is bounded for $i = 1, 2, \cdots, u$, then

(i) *There exists a minimax solution.*

(ii) *The class $\bar{C}$ has the property that for any δ not in $\bar{C}$ there exists an element δ^* of $\bar{C}$ such that $r(\theta, \delta^*) \leqq r(\theta, \delta)$ for all θ; i.e., $\bar{C}$ is essentially complete.*

As an illustration, we shall discuss the following simple example: The chance variables $X_1, \cdots, X_N$ are independently, normally, and identically distributed. The variance of the common normal distribution is assumed to be equal to 1 and the mean θ is unknown. Let a and b be two real numbers such that $a < b$. Also let H_1 be the hypothesis that $\theta \leqq a$, H_2 the hypothesis that $a < \theta < b$, and H_3 the hypothesis that $\theta \geqq b$. The space D^t consists of the elements $d_1{}^t$, $d_2{}^t$, and $d_3{}^t$, where $d_i{}^t$ denotes the decision to accept H_i.

Let ρ be a positive number $< (b - a)/2$. We put

$$
\begin{aligned}
&W_1(\theta) = 0 \quad \text{for } \theta < a + \rho, && = 1 \quad \text{for } \theta \geqq a + \rho \\
(5.84)\quad &W_2(\theta) = 0 \quad \text{for } a - \rho < \theta < b + \rho, && = 1 \quad \text{for all other } \theta \\
&W_3(\theta) = 0 \quad \text{for } \theta > b - \rho, && = 1 \quad \text{for } \theta \leqq b - \rho
\end{aligned}
$$

Let ξ be the a priori distribution in Ω, and ξ_x the a posteriori distribution after the sample $x = (x_1, \cdots, x_N)$ has been observed. Clearly

$$
\begin{aligned}
W_1(\xi_x) &= \int_{\theta \geqq a+\rho} d\xi_x(\theta) \\
(5.85)\qquad W_2(\xi_x) &= \int_{\theta \leqq a-\rho} d\xi_x(\theta) + \int_{\theta \geqq b+\rho} d\xi_x(\theta) \\
W_3(\xi_x) &= \int_{\theta \leqq b-\rho} d\xi_x(\theta)
\end{aligned}
$$

We shall now study the character of the set of x's for which

$$
(5.86)\qquad W_1(\xi_x) < W_2(\xi_x)
$$

The above inequality can be written as

$$
(5.87)\qquad \int_{a+\rho \leqq \theta < b+\rho} d\xi_x(\theta) < \int_{\theta \leqq a-\rho} d\xi_x(\theta)
$$

or

$$
(5.88)\qquad \int_{a+\rho \leqq \theta < b+\rho} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta) < \int_{\theta \leqq a-\rho} e^{N\bar{x}\theta - \frac{1}{2}N\theta^2}\, d\xi(\theta)
$$

where $\bar{x}$ is the arithmetic mean of the observations $x_1, \cdots, x_N$. We can easily verify that the set of values $\bar{x}$ for which the above inequality holds either is empty or is an open interval $(-\infty, c')$, where c' is a finite constant or ∞. We can also verify that, if c' is a finite constant,

$\bar{x} = c'$ is the only value of $\bar{x}$ for which the left-hand and right-hand members in (5.88) are equal. Thus, the set of all x satisfying $W_1(\xi_x) < W_2(\xi_x)$ is either the empty set, or the whole real axis, or there exists a finite constant c' such that $W_1(\xi_x) < W_2(\xi_x)$, $W_1(\xi_x) = W_2(\xi_x)$, or $W_1(\xi_x) > W_2(\xi_x)$, according to whether $\bar{x} < c'$, $\bar{x} = c'$, or $\bar{x} > c'$.

In a similar way we can show that the set of x's satisfying $W_1(\xi_x) < W_3(\xi_x)$ is either empty, or the whole real axis, or there exists a constant c'' such that $W_1(\xi_x) <$, or $=$, or $> W_3(\xi_x)$, according to whether $\bar{x} <$, or $=$, or $> c''$. Hence the set of all x's for which

$$W_1(\xi_x) < \text{Min}\,[W_2(\xi_x), W_3(\xi_x)] \tag{5.89}$$

is either empty, or the whole real axis, or there exists a finite constant c^* such that the relation $<$, or $=$, or $>$ holds between the two sides of (5.89), according to whether $\bar{x} <$, or $=$, or $> c^*$.

In a similar manner we find that the set of x's for which

$$W_3(\xi_x) < \text{Min}\,[W_1(\xi_x), W_2(\xi_x)] \tag{5.90}$$

is either empty, or the whole real axis, or there exists a finite constant c^{**} such that the relation $<$, $=$, or $>$ holds in (5.90) according to whether $\bar{x} >$, $=$, or $< c^{**}$.

The above results show that there exist two constants c_1 and c_2 (the improper values $-\infty$ and $+\infty$ being admitted) such that

$$W_1(\xi_x) < \text{Min}\,[W_2(\xi_x), W_3(\xi_x)] \tag{5.91}$$

when $\bar{x} < c_1$,

$$W_2(\xi_x) < \text{Min}\,[W_1(\xi_x), W_3(\xi_x)] \tag{5.92}$$

when $c_1 < \bar{x} < c_2$, and

$$W_3(\xi_x) < \text{Min}\,[W_1(\xi_x), W_2(\xi_x)] \tag{5.93}$$

when $\bar{x} > c_2$.

For any constants c_1 and c_2 ($c_1 \leqq c_2$), let δ^{c_1,c_2} denote the decision function for which $\delta_1{}^{c_1,c_2}(x) = 1$ when $\bar{x} < c_1$, $\delta_2{}^{c_1,c_2}(x) = 1$ when $c_1 \leqq \bar{x} \leqq c_2$, and $\delta_3{}^{c_1,c_2}(x) = 1$ when $\bar{x} > c_2$. It follows from the above results that any Bayes solution must be identical to δ^{c_1,c_2} for some values of c_1 and c_2 (the values $-\infty$ and $+\infty$ being admitted) except perhaps on a set of x's of Lebesgue measure zero. It is not difficult to prove that the converse is also true. We can even show that for

any c_1 and c_2 $(c_1 \leqq c_2)$ there exists an a priori distribution $\xi(\theta)$ such that $\xi(\theta)$ assigns the probability 1 to the set consisting of the three points $\theta = a - \rho$, $(a + b)/2$, $b + \rho$ and such that δ^{c_1,c_2} is a Bayes solution relative to $\xi(\theta)$. Thus the class C of all Bayes solutions coincides with the class of all δ^{c_1,c_2} corresponding to all possible values of c_1 and c_2.[14] Since the closure $\bar{C}$ of C is identical to C, Theorem 5.12 gives the following result: *For any decision function δ there exist two constants c_1 and c_2 $(c_1 \leqq c_2)$ such that $r(\theta, \delta^{c_1,c_2}) \leqq r(\theta, \delta)$ for all θ.*

5.2 Discussion of Some Specific Sequential Decision Problems

5.2.1 Introductory Remarks

The problem of choosing a terminal decision d^t after experimentation has been completed is essentially the same in the sequential as in the non-sequential case. Consider, for example, the problem of finding a Bayes solution relative to a given a priori probability measure ξ on Ω. Let ξ_x denote the a posteriori probability measure on Ω after experimentation has been terminated. A terminal decision $d_0{}^t$ is optimal if $W(\xi_x, d^t) = \int_\Omega W(F, d^t)\, d\xi_x$ takes its minimum value for $d^t = d_0{}^t$. This holds for sequential, as well as for non-sequential, decision problems. The main difficulty in constructing Bayes solutions in the sequential case lies in the problem of finding an optimum rule for carrying out the experimentation. No such problems arise, of course, in the treatment of the non-sequential case, where experimentation is carried out in a single stage by observing the values of the first N chance variables $X_1, \cdots, X_N$.

The purpose of the present Section 5.2 is to discuss in some detail a few specific problems which will serve as illustrations of the general theory and which are indicative of the nature of the difficulties that arise in the construction of optimum rules for carrying out the experimentation.

5.2.2 A Two-Sample Procedure for Testing the Mean of a Normal Distribution

We shall consider the following decision problem: The chance variables $X_1, X_2, \cdots$, etc., are known to be independently distributed with the same normal distribution. The variance of the common normal distribution is equal to 1, and the mean is known to be equal

[14] This characterization of the class of all Bayes solutions is contained as a special case in a more general result obtained by Sobel [48].

to one of the values $-\Delta$ and Δ, where Δ is a given positive number. Thus Ω consists of two elements F_1 and F_2 (say), where F_1 denotes the distribution of $X = \{X_i\}$ when the mean is $-\Delta$, and F_2 the distribution corresponding to the mean Δ. The space D^t is assumed to consist of the two elements $d_1{}^t$ and $d_2{}^t$, where $d_i{}^t$ denotes the decision to accept the hypothesis that F_i is the true distribution of X. Let

$$W(F_i, d_j{}^t) = 1 \quad \text{if } i \neq j, \qquad \text{and} = 0 \quad \text{if } i = j \tag{5.94}$$

We shall assume that the cost of experimentation is proportional to the number of observations. Let c denote the cost of a single observation. We shall assume, furthermore, that experimentation must be carried out in at most two stages; i.e., only decision functions δ are admitted according to which the probability is 1 that experimentation is carried out in at most two stages. Since the chance variables X_1, X_2, $\cdots$, etc., are independently and identically distributed, it is irrelevant which chance variables are observed, and we can assume without loss of generality that the first stage of the experiment consists of the observations on $X_1, \cdots, X_m$, and the second stage of the observations on $X_{m+1}, \cdots, X_{m+n}$, where m and n are non-negative integers. To simplify the problem further, we shall assume that m is a predetermined positive integer not chosen by the experimenter. Thus the experimenter's choice is restricted to the choice of n as a function of the observations $x_1, \cdots, x_m$.

A decision function δ can now be represented by a sequence of functions $\{\delta_i(x_1, \cdots, x_m)\}$ $(i = 0, 1, 2, \cdots,$ ad inf.) and a function $\delta_+(x_1, \cdots, x_{m+n})$ defined for all real values $x_1, \cdots, x_{m+n}$ and for any non-negative integer n. The functions δ_i are non-negative and $\sum_{i=0}^{\infty} \delta_i = 1$. Furthermore $\delta_+(x_1, \cdots, x_{m+n})$ can take only values between 0 and 1. On the basis of the above functions the decision procedure is carried out as follows: First we make the observations $x_1, \cdots, x_m$. To determine the size n of the second sample, we perform a chance experiment constructed so that the probability that the sample size i will be selected is equal to $\delta_i(x_1, \cdots, x_m)$ $(i = 0, 1, 2, \cdots,$ ad inf.). After both samples, consisting of the observations $x_1, \cdots, x_{m+n}$, have been drawn, the terminal decision is made with the help of a chance mechanism constructed so that the probability of accepting the hypothesis that Δ is the true mean is equal to $\delta_+(x_1, \cdots, x_{m+n})$.

We shall now study the nature of the Bayes solutions of this decision problem. Let ξ_i denote the a priori probability that F_i is true $(i = 1, 2)$. Also let ξ_{im} denote the a posteriori probability that F_i is

true ($i = 1, 2$) after the first m observations have been made; i.e.,

$$\xi_{1,m} = \frac{\xi_1 e^{-\Delta y_m}}{\xi_1 e^{-\Delta y_m} + \xi_2 e^{\Delta y_m}}$$
$$\xi_{2,m} = \frac{\xi_2 e^{\Delta y_m}}{\xi_1 e^{-\Delta y_m} + \xi_2 e^{\Delta y_m}} \tag{5.95}$$

where

$$y_u = x_1 + \cdots + x_u \tag{5.96}$$

for any integral value u. Let n denote the size of the second sample drawn. After both samples have been drawn, the a posteriori probability that F_i is true is given by $\xi_{i,m+n}$, where $\xi_{i,u}$ is defined by the expression we obtain from (5.95) when m is replaced by u.

A necessary condition for a decision function δ to be a Bayes solution is clearly the following: For any x (except perhaps on a set of measure zero),

$$\begin{aligned} \delta_+(x_1, \cdots, x_{m+n}) &= 1 \qquad \text{when } \xi_{2,m+n} > \tfrac{1}{2} \\ \delta_+(x_1, \cdots, x_{m+n}) &= 0 \qquad \text{when } \xi_{2,m+n} < \tfrac{1}{2} \end{aligned} \tag{5.97}$$

Since the set of samples $(x_1, \cdots, x_{m+n})$ for which $\xi_{2,m+n} = \frac{1}{2}$ is of measure zero, the above condition determines $\delta_+(x_1, \cdots, x_{m+n})$ uniquely (except on a set of measure zero). Thus the problem of finding a Bayes solution is reduced to the problem of a proper choice of the functions $\delta_i(x_1, \cdots, x_m)$ ($i = 0, 1, \cdots$, ad inf.). To deal with the latter problem, we shall study the conditional risk when $x_1, \cdots, x_m$ and n are given and $\delta_+(x_1, \cdots, x_{m+n})$ satisfies (5.97).

For given $x_1, \cdots, x_m$ and n, let $\alpha(x_1, \cdots, x_m, n)$ be the conditional probability that we shall accept F_2 when F_1 is true, and $\beta(x_1, \cdots, x_m, n)$ the conditional probability that we shall accept F_1 when F_2 is true. Thus $\alpha(x_1, \cdots, x_m, n)$ is equal to the conditional probability of the inequality

$$\xi_2 e^{\Delta y_{m+n}} > \xi_1 e^{-\Delta y_{m+n}} \tag{5.98}$$

when F_1 is the true distribution, and $1 - \beta(x_1, \cdots, x_m, n)$ is equal to the conditional probability that (5.98) holds when F_2 is true. The above inequality can be written as

$$e^{2\Delta z_n} > \frac{\xi_{1,m}}{\xi_{2,m}} = h_m \quad \text{(say)} \tag{5.99}$$

where

$$z_n = y_{m+n} - y_m \tag{5.100}$$

If $n \geqq 1$, the conditional probability that (5.99) holds can easily be expressed in terms of the Gaussian function

$$G(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{1}{2}u^2}\, du \tag{5.101}$$

One can readily obtain the following results:

$$\alpha(x_1, \cdots, x_m, n) = G\left[\frac{\log h_m}{2\Delta\sqrt{n}} + \sqrt{n}\Delta\right] \qquad (n \geqq 1) \tag{5.102}$$

and

$$1 - \beta(x_1, \cdots, x_m, n) = G\left[\frac{\log h_m}{2\Delta\sqrt{n}} - \sqrt{n}\Delta\right] \qquad (n \geqq 1) \tag{5.103}$$

For $n = 0$, we have

$$\alpha(x_1, \cdots, x_m, 0) = 1 - \beta(x_1, \cdots, x_m, 0) = 0 \quad \text{when } h_m > 1$$
$$= 1 \quad \text{when } h_m < 1 \tag{5.104}$$

Since $\xi_{i,m}$ is the a posteriori probability of F_i $(i = 1, 2)$ when $x_1, \cdots, x_m$ are given, the a posteriori risk when $x_1, \cdots, x_m$ and n are given is equal to

$$r(x_1, \cdots, x_m, n)$$
$$= \xi_{1m}\alpha(x_1, \cdots, x_m, n) + \xi_{2m}\beta(x_1, \cdots, x_m, n) + c(m + n) \tag{5.105}$$

For any given values $x_1, \cdots, x_m$, we shall be interested in values of n for which $r(x_1, \cdots, x_m, n)$ takes its minimum value. Clearly the minimum of $r(x_1, \cdots, x_m, n)$ with respect to n can be taken only for values $n \leqq 1/c$. For any given sample $(x_1, \cdots, x_m)$, let $N(x_1, \cdots, x_m)$ denote the set of all integral values n for which $r(x_1, \cdots, x_m, n)$ takes its minimum value. Thus $N(x_1, \cdots, x_m)$ is a subset of the set of all non-negative integers not exceeding $1/c$.

A necessary and sufficient condition for a decision function δ to be a Bayes solution relative to a given a priori distribution $\xi = (\xi_1, \xi_2)$ is given as follows: $\delta_+(x_1, \cdots, x_{m+n})$ *satisfies* (5.97), *and* $\delta_i(x_1, \cdots, x_m) = 0$ *for any* i *that is not an element of* $N(x_1, \cdots, x_m)$ (*except perhaps on a set of Lebesgue measure zero*). If $N(x_1, \cdots, x_m)$ consists of a single element for each sample $(x_1, \cdots, x_m)$, then there is (essentially) only one Bayes solution.

To compute the set $N(x_1, \cdots, x_m)$, it is helpful to consider n as a continuous variable; i.e., n is allowed to take any non-negative real value. For non-integral values n, we define $\alpha(x_1, \cdots, x_m, n)$ and $\beta(x_1, \cdots, x_m, n)$ formally by the equations (5.102) and (5.103). The

partial derivative of $r(x_1, \cdots, x_m, n)$ with respect to n is then given by

$$\frac{\partial r(x_1, \cdots, x_m, n)}{\partial n} \tag{5.106}$$

$$= \xi_{1m}\left\{\frac{-1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log h_m}{2\Delta\sqrt{n}}+\Delta\sqrt{n}\right)^2}\left[\frac{-\log h_m}{4\Delta n^{3/2}}+\frac{1}{2}\frac{\Delta}{\sqrt{n}}\right]\right\}$$

$$-\xi_{2m}\left\{-\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log h_m}{2\Delta\sqrt{n}}-\Delta\sqrt{n}\right)^2}\left[\frac{-\log h_m}{4\Delta n^{3/2}}-\frac{1}{2}\frac{\Delta}{\sqrt{n}}\right]\right\}+c$$

The set $N(x_1, \cdots, x_m)$ can be established by studying the roots of the equation

$$\frac{\partial r(x_1, \cdots, x_m, n)}{\partial n} = 0 \tag{5.107}$$

in n.

It is of interest to note that $r(x_1, \cdots, x_m, n)$ is a function of $y_m = x_1 + \cdots + x_m$ and n only. This follows immediately from equations (5.102) to (5.105). Hence the set $N(x_1, \cdots, x_m)$ depends only on $y_m = x_1 + \cdots + x_m$. This fact greatly facilitates the tabulation of $N(x_1, \cdots, x_m)$. Since for any n the value of $r(x_1, \cdots, x_m, n)$ can easily be computed on the basis of formulas (5.102) to (5.105), the set $N(x_1, \cdots, x_m)$ can be established by trial and error.

One can easily verify that Assumptions 3.1 to 3.7 of Chapter 3 are fulfilled for the decision problem under consideration here. Thus all the results obtained in Chapter 3 can be applied to this problem. In particular, the following statements are true: (i) *The class of all Bayes solutions relative to all possible a priori distributions ξ is a complete class of decision functions;* (ii) *there exists a least favorable a priori distribution ξ;* (iii) *there exists a minimax solution;* (iv) *a minimax solution is also a Bayes solution relative to any least favorable a priori distribution ξ.*

We shall now show that $\xi^0 = (\frac{1}{2}, \frac{1}{2})$ is a least favorable a priori distribution. It follows from reasons of symmetry that when ξ^0 is the a priori distribution we have $N(x_1, \cdots, x_m) = N(-x_1, \cdots, -x_m)$. Hence there exists a Bayes solution δ^0 relative to ξ^0 such that $\delta_i{}^0(x_1, \cdots, x_m) = \delta_i{}^0(-x_1, \cdots, -x_m)$. Furthermore for any Bayes solution δ^0 relative to ξ^0 we have $\delta_+(x_1, \cdots, x_{m+n}) = 1$ if $x_1 + \cdots + x_{m+n} > 0$, and $= 0$ when $x_1 + \cdots + x_{m+n} < 0$. It follows from symmetry considerations that for any Bayes solution δ^0 relative to ξ^0 for which $\delta_i(x_1, \cdots, x_m) = \delta_i(-x_1, \cdots, -x_m)$, we have

$$r(F_1, \delta^0) = r(F_2, \delta^0) \tag{5.108}$$

The above equation shows that $\xi^0 = (\frac{1}{2}, \frac{1}{2})$ must be a least favorable

a priori distribution and any Bayes solution δ^0 relative to ξ^0 satisfying the condition $\delta_i(x_1, \cdots, x_m) = \delta_i(-x_1, \cdots, -x_m)$ $(i = 0, 1, 2, \cdots,$ ad inf.) must be a minimax solution.

It was remarked before that $r(x_1, \cdots, x_m, n)$ depends only on $y_m = x_1 + \cdots + x_m$ and n. Thus we may write $r(y_m, n)$ instead of $r(x_1, \cdots, x_m, n)$. Let

$$\rho(y_m) = \mathrm{Min}_n\, r(y_m, n) \tag{5.109}$$

Furthermore let $\psi_i(m)$ denote the expected value of $\rho(y_m)$ when F_i is true $(i = 1, 2)$; i.e.,

$$\psi_1(m) = \frac{1}{\sqrt{2\pi}\sqrt{m}} \int_{-\infty}^{\infty} \rho(y_m) e^{-\frac{1}{2m}(y_m + m\Delta)^2}\, dy_m \tag{5.110}$$

and

$$\psi_2(m) = \frac{1}{\sqrt{2\pi}\sqrt{m}} \int_{-\infty}^{\infty} \rho(y_m) e^{-\frac{1}{2m}(y_m - m\Delta)^2}\, dy_m \tag{5.111}$$

Then for any a priori distribution $\xi = (\xi_1, \xi_2)$ we have

$$\mathrm{Min}_\delta\, r(\xi, \delta) = \xi_1\psi_1(m) + \xi_2\psi_2(m) \tag{5.112}$$

If the value of m can be chosen by the experimenter, an optimum choice would be a value m for which the right-hand side of (5.112) becomes a minimum. The functions $\psi_1(m)$ and $\psi_2(m)$ are, however, difficult to compute. A crude approximation to $\psi_i(m)$, but perhaps sufficient for some practical purposes, may be obtained from $\rho(y_m)$ by replacing y_m by its expected value under F_i; i.e.,

$$\psi_1(m) \sim \rho(-m\Delta) \quad \text{and} \quad \psi_2(m) \sim \rho(m\Delta) \tag{5.113}$$

5.2.3 A Sequential Procedure for Testing the Means of a Pair of Binomial Distributions

In this section we shall discuss the following decision problem: Let $X_1, X_2, \cdots$, etc., be independently distributed chance variables where each chance variable X_i can take only the values 0 and 1. The probability that $X_i = 1$ is equal to p when i is odd, and equal to p^* when i is even. The constants p and p^* are unknown, but it is known that (p, p^*) is equal to either (p_1, p_1^*) or (p_2, p_2^*), where p_1, p_2, p_1^*, and p_2^* are given positive numbers < 1. We shall assume that $p_1 \neq p_2$ and $p_1^* \neq p_2^*$. Let F_i represent the distribution of the sequence $X = \{X_i\}$ when $(p, p^*) = (p_i, p_i^*)$ $(i = 1, 2)$. Thus the space Ω consists of the two elements F_1 and F_2. The space D^t is assumed to consist of the two elements $d_1{}^t$ and $d_2{}^t$, where $d_i{}^t$ denotes the terminal decision to accept the hypothesis that F_i is the true distribution of X. We put

$$W(F_i, d_j{}^t) = 1 \quad \text{if } i \neq j, \qquad \text{and} \qquad = 0 \quad \text{if } i = j \tag{5.114}$$

The cost of experimentation is assumed to be proportional to the total number of observations made and to be independent of anything else. Thus, if n is the total number of observations made, the cost is cn, where c denotes the cost of a single observation.

It will be convenient to use the symbol Y_i to denote the chance variable X_{2i-1} ($i = 1, 2, \cdots$, ad inf.), and Z_i to denote the chance variable X_{2i} ($i = 1, 2, \cdots$, ad inf.). Similarly y_i will stand for x_{2i-1}, and z_i for x_{2i}. Since the cost of experimentation does not depend on the number of stages in which the experiment is carried out, we can restrict ourselves to decision functions δ according to which each stage of the experiment consists of precisely one observation. Since the Y chance variables, as well as the Z chance variables, have a common distribution, we may impose the following further restriction on δ: Whenever a y-observation is made, we observe the value of Y_i with the smallest index i that has not yet been observed; and whenever we make a z-observation, we observe the value of Z_i with the smallest index i that has not yet been observed. In view of the above restrictions, a decision function δ can be represented by four functions $\delta_i(y_1, \cdots, y_m; z_1, \cdots, z_n)$ ($i = 1, 2, 3, 4$) satisfying the conditions

$$\delta_i(y_1, \cdots, y_m; z_1, \cdots, z_n) \geqq 0$$

$$\sum_{i=1}^{4} \delta_i(y_1, \cdots, y_m; z_1, \cdots, z_n) = 1 \tag{5.115}$$

Here $\delta_i(y_1, \cdots, y_m; z_1, \cdots, z_n)$ denotes the probability of making the terminal decision d_i^t ($i = 1, 2$) when the sample $y_1, \cdots, y_m; z_1, \cdots, z_n$ has been observed; $\delta_3(y_1, \cdots, y_m; z_1, \cdots, z_n)$ is the probability of continuing experimentation by observing the value of Y_{m+1}; and $\delta_4(y_1, \cdots, y_m; z_1, \cdots, z_n)$ is the probability of continuing experimentation by observing the value of Z_{n+1}. The functions δ_i are defined also for $m = n = 0$. If $m = n = 0$, δ_i is equal to the probability of the corresponding action before the start of experimentation.

The problem under consideration here is somewhat different from the type of problems treated in Chapter 4, since in Chapter 4 we assumed that all chance variables X_i ($i = 1, 2, \cdots$, ad inf.) have the same distribution. Nevertheless, most of the results in Chapter 4 will be applicable to the present case, as will be indicated later.

Any probability distribution in Ω can be represented by a number ξ between zero and 1, where ξ denotes the probability that F_1 is true. For any ξ, let $\xi_{z_1, \cdots, z_n}^{y_1, \cdots, y_m}$ denote the a posteriori probability distribution in Ω (the a posteriori probability that F_1 is true) when ξ is the a priori probability distribution and the sample $(y_1, \cdots, y_m; z_1, \cdots, z_n)$ has

been observed. If $n = 0$, the above symbol reduces to $\xi^{y_1,\cdots,y_m}$. Similarly, if $m = 0$, the above symbol reduces to $\xi_{z_1,\cdots,z_n}$. If $m = n = 0$, the symbol $\xi^{y_1,\cdots,y_m}_{z_1,\cdots,z_n}$ reduces to ξ. Clearly

$$(5.116)\quad \xi^{y_1,\cdots,y_m}_{z_1,\cdots,z_n} = \frac{\xi p_1^{\sum_1^m y_i}(1-p_1)^{m-\sum_1^m y_i}(p_1^*)^{\sum_1^n z_i}(1-p_1^*)^{n-\sum_1^n z_i}}{\left[\begin{array}{l}\xi p_1^{\sum_1^m y_i}(1-p_1)^{m-\sum_1^m y_i}(p_1^*)^{\sum_1^n z_i}(1-p_1^*)^{n-\sum_1^n z_i} \\ \quad + (1-\xi)p_2^{\sum_1^m y_i}(1-p_2)^{m-\sum_1^m y_i}(p_2^*)^{\sum_1^n z_i}(1-p_2^*)^{n-\sum_1^n z_i}\end{array}\right]}$$

For any non-negative integer k, let δ^k denote a decision function such that the probability is zero that more than k observations will be made when δ^k is adopted. As in Chapter 4, we define

$$(5.117)\quad \rho_k(\xi) = \mathrm{Inf}_{\delta^k} r(\xi, \delta^k) \quad (k = 0, 1, 2, \cdots, \text{ad inf.})$$

Clearly

$$(5.118)\quad \rho_0(\xi) = \mathrm{Min}\ (\xi, 1-\xi)$$

The following recursion formula holds:

$$(5.119)\quad \rho_{k+1}(\xi) = \mathrm{Min}\ [\rho_0(\xi), a_0(\xi)\rho_k(\xi^0) + a_1(\xi)\rho_k(\xi^1) + c, b_0(\xi)\rho_k(\xi_0) + b_1(\xi)\rho_k(\xi_1) + c]$$

where

$$(5.120)\quad \begin{aligned} a_0(\xi) &= \xi(1-p_1) + (1-\xi)(1-p_2) \\ a_1(\xi) &= \xi p_1 + (1-\xi)p_2 \end{aligned}$$

and

$$(5.121)\quad \begin{aligned} b_0(\xi) &= \xi(1-p_1^*) + (1-\xi)(1-p_2^*) \\ b_1(\xi) &= \xi p_1^* + (1-\xi)p_2^* \end{aligned}$$

The proof of the above recursion formula is omitted, since it is essentially the same as that of the corresponding recursion formulas given in Chapter 4 (see Theorem 4.1).

The relation

$$(5.122)\quad \lim_{k=\infty} \rho_k(\xi) = \rho(\xi) = \mathrm{Inf}_\delta r(\xi, \delta)$$

can be proved in exactly the same way as the corresponding relation in Chapter 4 was proved; see equation (4.21).

The functions $\rho_0(\xi)$, $\rho_1(\xi)$, $\rho_2(\xi)$, $\cdots$, $\rho_{k_0}(\xi)$ can be used to give a complete characterization of Bayes solutions when only decision functions δ are admitted for which the probability is zero that the number of observations needed for the experiment will exceed a prescribed number k_0. For this purpose, we shall define three subsets, $S_{u,1}$, $S_{u,2}$, and $S_{u,3}$ of the interval [0, 1], depending on a parameter u which can take only positive integral values. $S_{u,1}$ is defined as the set of all values ξ for which

$$\rho_0(\xi) > \rho_u(\xi) \tag{5.123}$$

$S_{u,2}$ is the set of values ξ for which

$$a_0(\xi)\rho_{u-1}(\xi^0) + a_1(\xi)\rho_{u-1}(\xi^1) + c > \text{Min}\,[\rho_0(\xi), b_0(\xi)\rho_{u-1}(\xi_0) + b_1(\xi)\rho_{u-1}(\xi_1) + c] \tag{5.124}$$

$S_{u,3}$ is the set of all values ξ for which

$$b_0(\xi)\rho_{u-1}(\xi_0) + b_1(\xi)\rho_{u-1}(\xi_1) + c > \text{Min}\,[\rho_0(\xi), a_0(\xi)\rho_{u-1}(\xi^0) + a_1(\xi)\rho_{u-1}(\xi^1) + c] \tag{5.125}$$

A decision function δ, subject to the restriction that the probability is zero that the number of observations will exceed k_0, is a Bayes solution relative to the a priori distribution ξ if and only if the following five conditions are fulfilled:

$$\delta_1(y_1, \cdots, y_m; z_1, \cdots, z_n) = 0 \tag{5.126}$$

if $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m} < \frac{1}{2}$,

$$\delta_2(y_1, \cdots, y_m; z_1, \cdots, z_n) = 0 \tag{5.127}$$

if $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m} > \frac{1}{2}$,

$$\delta_1(y_1, \cdots, y_m; z_1, \cdots, z_n) = \delta_2(y_1, \cdots, y_m; z_1, \cdots, z_n) = 0 \tag{5.128}$$

if $m + n < k_0$ and $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is an element of $S_{k_0-m-n,1}$,

$$\delta_3(y_1, \cdots, y_m; z_1, \cdots, z_n) = 0 \tag{5.129}$$

if $m + n = k_0$, or if $m + n < k_0$ and $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is an element of $S_{k_0-m-n,2}$,

$$\delta_4(y_1, \cdots, y_m; z_1, \cdots, z_n) = 0 \tag{5.130}$$

if $m + n = k_0$, or if $m + n < k_0$ and $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is an element of $S_{k_0-m-n,3}$.

The above five conditions are satisfied for the following decision rule: Continue experimentation as long (and only as long) as $m + n < k_0$ and $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is an element of $S_{k_0-m-n,1}$. If the sample $(y_1, \cdots, y_m; z_1, \cdots, z_n)$ has already been obtained and the application of the above

rule requires taking an additional observation, take a y-observation when $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ lies in $S_{k_0-m-n,3}$, and a z-observation otherwise. If experimentation is terminated with the sample $(y_1, \cdots, y_m; z_1, \cdots, z_n)$, decide for $d_1{}^t$ when $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m} \geqq \frac{1}{2}$, and for $d_2{}^t$ when $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m} < \frac{1}{2}$.

Let S_i denote the set of values ξ to which $S_{u,i}$ reduces if u is replaced by ∞ and $\rho_\infty(\xi)$ by $\rho(\xi)$ $(i = 1, 2, 3)$. A characterization of Bayes solutions with no restrictions on the number of observations can be given in terms of the sets S_i. A decision function δ is a Bayes solution relative to ξ if and only if (5.126) to (5.130) are satisfied when k_0 is replaced by ∞ and $S_{\infty,i}$ by S_i. A decision rule satisfying these conditions may be given as follows: Continue experimentation as long (and only as long) as $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is an element of S_1. If $(y_1, \cdots, y_m; z_1, \cdots, z_n)$ has already been observed and if $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is in S_1, take a y-observation when $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is in S_3, and a z-observation otherwise. If the sample $(y_1, \cdots, y_m; z_1, \cdots, z_n)$ has already been obtained and $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m}$ is not in S_1, decide for $d_1{}^t$ when $\xi_{z_1,\cdots,z_n}^{y_1,\cdots,y_m} \geqq \frac{1}{2}$, and for $d_2{}^t$ otherwise.

Let $S_i{}^*$ denote the complement of S_i; i.e., $S_i{}^*$ consists of all values ξ which do not belong to S_i. The intersection of $S_1{}^*$ with the interval $[0, \frac{1}{2}]$ is the set that was denoted in Chapter 4 by $C_{d_1{}^t}$, and the intersection of $S_1{}^*$ with $[\frac{1}{2}, 1]$ is the set that was denoted in Chapter 4 by $C_{d_1{}^t}$. The proof that the sets $C_{d_1{}^t}$ and $C_{d_2{}^t}$ are closed and convex, given in Chapter 4, applies without modification to the present case. Thus $C_{d_1{}^t}$ and $C_{d_2{}^t}$ are closed intervals. Clearly $C_{d_1{}^t}$ contains the point $\xi = 1$, and $C_{d_2{}^t}$ contains the point $\xi = 0$. Let a be the upper endpoint of $C_{d_2{}^t}$, and b the lower endpoint of $C_{d_1{}^t}$. Clearly $S_1{}^*$ is the set-theoretical sum of the intervals $[0, a]$ and $[b, 1]$. Thus S_1 is the open interval (a, b).

It can be shown that the intersection S_2S_3 of the sets S_2 and S_3 is precisely equal to the set-theoretical sum of the half-open intervals $[0, a)$ and $(b, 1]$. The proof of this is omitted because it is essentially the same as that of Theorem 4.9 in Chapter 4.

By using the above results concerning the nature of the sets S_1, S_2, and S_3, the characterization of Bayes solutions can be given in the following form: *A decision function δ is a Bayes solution if and only if it satisfies the following conditions:* (i) *If the a posteriori probability of F_1 is $< a$ at some stage of the experiment,*[15] *experimentation is stopped and the terminal decision $d_2{}^t$ is made.* (ii) *If the a posteriori probability of F_1 is $> b$, experimentation is stopped and the terminal decision $d_1{}^t$ is made.* (iii) *If the a posteriori probability of F_1 is $> a$ and $< b$, an additional observation is made.* (iv) *If an additional observation is made, it*

[15] The a posteriori probability of F_1 is to be replaced by the a priori probability of F_1 when experimentation has not yet started.

must be a y-observation if the a posteriori probability of F_1 is a point in S_3, and a z-observation if this a posteriori probability is a point in S_2. (v) *If experimentation is terminated when the a posteriori probability of F_1 is equal to a (b) and if $a < \frac{1}{2}$ ($b > \frac{1}{2}$), the terminal decision $d_2{}^t(d_1{}^t)$ is made.*

Let S'_i be the intersection of S_i with the open interval (a, b) $(i = 2, 3)$. Since no point ξ in (a, b) belongs to the intersection of S_2 and S_3, the sets S'_2 and S'_3 are disjoint. The set S'_2 consists precisely of those points ξ in (a, b) for which

$$a_0(\xi)\rho(\xi^0) + a_1(\xi)\rho(\xi^1) > b_0(\xi)\rho(\xi_0) + b_1(\xi)\rho(\xi_1) \tag{5.131}$$

and the set S'_3 consists of the points ξ in (a, b) for which

$$a_0(\xi)\rho(\xi^0) + a_1(\xi)\rho(\xi^1) < b_0(\xi)\rho(\xi_0) + b_1(\xi)\rho(\xi_1) \tag{5.132}$$

The nature of the sets defined by the inequalities (5.131) and (5.132) has not been studied. It is not unlikely that these sets have a simple structure; perhaps they are frequently intervals.

Clearly Assumptions 3.1 to 3.7 of Chapter 3 are fulfilled for the decision problem under consideration here. Thus all results obtained in Chapter 3 are applicable to this case. In particular, the following statements hold: (i) The class of all Bayes solutions is a complete class of decision functions; (ii) there exists a value ξ' such that $\xi = \xi'$ is a least favorable a priori distribution; (iii) a minimax solution exists and any minimax solution is a Bayes solution relative to any least favorable a priori distribution.

The foregoing results can easily be generalized in two directions: (1) Instead of assuming that each Y-variable and each Z-variable can take only the values 0 and 1, we may work with any general (absolutely continuous or discrete) common distribution for the Y-variables, and any general common distribution for the Z-variables; (2) instead of assuming that the sequence $\{X_i\}$ can be split into two subsequences such that the chance variables belonging to the same subsequence have a common distribution, the more general case can be treated where $\{X_i\}$ can be split into a finite number of disjoint subsequences such that the chance variables belonging to the same subsequence have a common distribution.

5.2.4 Discussion of a Decision Problem when Ω Consists of Three Rectangular Distributions

As an illustration of the various ideas and notions of the general decision theory, we shall discuss here a rather simple decision problem.

The chance variables $X_1, X_2, \cdots$, etc., are assumed to be independently and identically distributed. The common distribution is known to be a rectangular distribution with unit range. The midpoint θ of the range is unknown, but it is known that it is equal to one of the values: $-\frac{1}{4}$, 0, and $\frac{1}{4}$. Thus in this problem Ω consists of three elements. The space D^t of terminal decisions is assumed to consist of three elements $d_1{}^t$, $d_2{}^t$, and $d_3{}^t$, where $d_1{}^t$ denotes the decision to reject the hypothesis H_1 that $\theta = -\frac{1}{4}$, $d_2{}^t$ denotes the decision to reject the hypothesis H_2 that $\theta = 0$, and $d_3{}^t$ denotes the decision to reject the hypothesis H_3 that $\theta = \frac{1}{4}$. Let $W(F_i, d_j{}^t) = 1$ if $i = j$, and $= 0$ if $i \neq j$, where F_i denotes the distribution of $X = \{X_j\}$ when H_i is true. In other words, the loss due to the terminal decision $d_i{}^t$ is 1 if H_i is true, and 0 if H_i is not true. The cost of experimentation is assumed to be proportional to the number of observations. Let c be the cost of a single observation. We shall assume that $0 < c < \frac{1}{2}$.

The above decision problem is a special case of the general decision problem treated in Chapter 4. Thus we shall use the terminology and notation adopted in Chapter 4. An a priori distribution in Ω is given by a vector $\xi = (\xi^1, \xi^2, \xi^3)$, where ξ^i denotes the a priori probability that H_i is true ($i = 1, 2, 3$).

Let $x = (x_1, \cdots, x_m)$ be a sample of m observations (x_i is the observed value of X_i), and let ξ_x denote the a posteriori distribution when ξ is the a priori distribution and $x = (x_1, \cdots, x_m)$ is the observed sample. Clearly $\xi_x = \xi$ if $-\frac{1}{4} \leqq x_i \leqq \frac{1}{4}$ for $i = 1, \cdots, m$. If Min $(x_1, \cdots, x_m) < -\frac{1}{4}$, then $\xi_x{}^3 = 0$. If Max $(x_1, \cdots, x_m) > \frac{1}{4}$, then $\xi_x{}^1 = 0$. Thus, if Min $(x_1, \cdots, x_m) < -\frac{1}{4}$ or Max $(x_1, \cdots, x_m) > \frac{1}{4}$, we can make a terminal decision without any (a posteriori) risk. The probability that an observation will lie outside the interval $[-\frac{1}{4}, \frac{1}{4}]$ is equal to $\frac{1}{2}$ under each hypothesis H_i ($i = 1, 2, 3$). Hence for any a priori distribution ξ the probability is equal to $\frac{1}{2}$ that an observation will fall outside $[-\frac{1}{4}, \frac{1}{4}]$. From this it follows that if experimentation is continued until we obtain an observation outside $[-\frac{1}{4}, \frac{1}{4}]$, the expected number of observations is equal to 2. Hence

$$\rho(\xi) \leqq 2c \tag{5.133}$$

where $\rho(\xi) = \mathrm{Inf}_\delta\, r(\xi, \delta)$ [see Section 4.1.1]. The minimum risk $\rho_0(\xi)$—see equation (4.8) in Section 4.1.1—when a terminal decision is to be made without any experimentation is given by

$$\rho_0(\xi) = \mathrm{Min}\ (\xi^1, \xi^2, \xi^3) \tag{5.134}$$

Since the a posteriori distribution ξ_x coincides with the a priori prob-

ability distribution ξ as long as no observation falls outside the interval $[-\frac{1}{4}, \frac{1}{4}]$, it is clear that

$$\rho(\xi) = \text{Min}\,[\rho_0(\xi), 2c] = \text{Min}\,(\xi^1, \xi^2, \xi^3, 2c) \tag{5.135}$$

Bayes solutions can easily be constructed with the help of the functions $\rho_0(\xi)$ and $\rho(\xi)$. We have to consider the following three cases.

I. *Min* $(\xi^1, \xi^2, \xi^3) > 2c$. In this case a Bayes solution is given as follows: We take observations until we obtain one that lies outside the interval $[-\frac{1}{4}, \frac{1}{4}]$. If the last observation x_n is $< -\frac{1}{4}$ and $\geqq -\frac{1}{2}$, we choose the terminal decision $d_3{}^t$. If $x_n < -\frac{1}{2}$ we may choose between $d_2{}^t$ and $d_3{}^t$ at random. If $x_n > \frac{1}{4}$ and $\leqq \frac{1}{2}$, we choose $d_1{}^t$. If $x_n > \frac{1}{2}$, we may choose between $d_1{}^t$ and $d_2{}^t$ at random. Using the notation introduced in Section 4.1.1, we can express this as follows: $\delta(i+1 \mid x_1, \cdots, x_i) = 1$ if $-\frac{1}{4} \leqq \text{Min}\,(x_1, \cdots, x_i) \leqq \text{Max}\,(x_1, \cdots, x_i) \leqq \frac{1}{4}$, and $\delta(i+1 \mid x_1, \cdots, x_i) = 0$ otherwise. $\delta(d_i{}^t \mid x_1, \cdots, x_n) = 0$ if the observations $x_1, \cdots, x_n$ fall inside the range corresponding to H_i.

II. *Min* $(\xi^1, \xi^2, \xi^3) = 2c$. For a decision function δ to be a Bayes solution it is necessary and sufficient that the following conditions be fulfilled (except perhaps on a set of samples whose probability measure is zero according to H_1, H_2, and H_3): (1) $\delta(d_i{}^t \mid 0) = 0$ for any i for which $\xi^i > \text{Min}\,(\xi^1, \xi^2, \xi^3)$; (2) $\delta(d_i{}^t \mid x_1, \cdots, x_r) = 0$ if $\xi^i > \text{Min}\,(\xi^1, \xi^2, \xi^3)$ and all the observations $x_1, \cdots, x_r$ are inside the interval $[-\frac{1}{4}, \frac{1}{4}]$; (3) $\delta(r+1 \mid x_1, \cdots, x_r) = 0$ if $x_1, \cdots, x_{r-1}$ are inside $[-\frac{1}{4}, \frac{1}{4}]$ and x_r is outside $[-\frac{1}{4}, \frac{1}{4}]$; (4) $\delta(d_i{}^t \mid x_1, \cdots, x_n) = 0$ if the observations $x_1, \cdots, x_{n-1}$ are inside $[-\frac{1}{4}, \frac{1}{4}]$ and x_n is outside $[-\frac{1}{4}, \frac{1}{4}]$ but inside the range corresponding to H_i.

III. *Min* $(\xi^1, \xi^2, \xi^3) < 2c$. In this case a necessary and sufficient condition for a decision function δ to be a Bayes solution is that $\sum_{i=1}^{3} \delta(d_i{}^t \mid 0) = 1$ and $\delta(d_j{}^t \mid 0) = 0$ for any j for which $\xi^j > \text{Min}\,(\xi^1, \xi^2, \xi^3)$.

Let δ_0 be the decision function determined as follows: $\delta_0(1 \mid 0) = 1$. $\delta(i+1 \mid x_1, \cdots, x_i) = 1$ if $-\frac{1}{4} \leqq x_j \leqq \frac{1}{4}$ for $j = 1, \cdots, i$. $\delta(d_i{}^t \mid x_1, \cdots, x_n) = 1$ if the observations $x_1, \cdots, x_{n-1}$ are inside the interval $[-\frac{1}{4}, \frac{1}{4}]$, x_n is outside $[-\frac{1}{4}, \frac{1}{4}]$, and i is the smallest integer such that the range corresponding to H_i does not contain x_n. Clearly $r(\theta, \delta_0) = 2c$ for $\theta = -\frac{1}{4}, 0, \frac{1}{4}$. Hence, if we can show that δ_0 is a Bayes solution relative to some a priori distribution ξ, δ_0 is a minimax solution. Let $c < \frac{1}{6}$. Then δ_0 is a Bayes solution relative to $\xi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Thus δ_0 is a minimax solution when $c < \frac{1}{6}$.

Let C_0 be the class of all decision functions δ which satisfy the following two conditions (except perhaps on a set of samples whose probability measure is zero according to H_1, H_2, and H_3):

(i) If $x_1, \cdots, x_{n-1}$ are in $[-\frac{1}{4}, \frac{1}{4}]$, x_n is outside $[-\frac{1}{4}, \frac{1}{4}]$, and if $\delta(1 \mid 0)\delta(2 \mid x_1) \cdots \delta(n \mid x_1, \cdots, x_{n-1}) > 0$, then $\delta(n+1 \mid x_1, \cdots, x_n) = 0$ and $\delta(d_i^t \mid x_1, \cdots, x_n) = 0$ for any i for which the range corresponding to H_i contains x_n.

(ii) There exists a positive integer $i \leqq 3$ such that $\delta(d_i^t \mid x_1, \cdots, x_r) = 0$ for any sample $x_1, \cdots, x_r$ for which $-\frac{1}{4} \leqq x_j \leqq \frac{1}{4}$ for $j = 1, \cdots, r$ and for which $\delta(1 \mid 0)\delta(2 \mid x_1) \cdots \delta(r \mid x_1, \cdots, x_{r-1}) > 0$.

We shall show that C_0 is a minimal complete class of decision functions if $0 < c < \frac{1}{6}$. First we show that, if δ is a Bayes solution relative to some a priori distribution ξ, δ is a member of C_0. Clearly δ must satisfy (i). Since $c < \frac{1}{6}$, there exists a positive integer $i \leqq 3$ such that $\xi^i > 2c$. A Bayes solution δ must satisfy condition (ii) for any i for which $\xi^i > 2c$. We shall now show the converse. Let δ be any member of C_0, i.e., any decision function which satisfies (i) and (ii). Suppose that (ii) is satisfied for $i = i_0$. Let $\xi_c = (\xi_c^1, \xi_c^2, \xi_c^3)$ be the a priori distribution given as follows: $\xi_c^{i_0} = 1 - 4c$ and $\xi_c^j = 2c$ for $j \neq i_0$. Clearly δ is a Bayes solution relative to ξ_c. Furthermore δ must be an admissible decision function, since all components of ξ_c are positive. Hence C_0 is identical with the class of all Bayes solutions and each member of C_0 is an admissible decision function. It then follows from Theorem 3.20 that C_0 is a minimal complete class of decision functions.

5.2.5 Sequential Point Estimation of the Mean of a Rectangular Distribution with Unit Range

In this section we shall discuss the following problem: Let X_1, X_2, $\cdots$, etc., be independently and identically distributed chance variables. The common distribution is known to be a rectangular distribution with unit range, but the mean θ is unknown (may take any real value). Thus Ω is a one-parameter family of distribution functions. The problem is to set up a point estimate for θ. For any real value θ^*, let $d_{\theta^*}{}^t$ denote the terminal decision to estimate the unknown mean θ by the value θ^*. Thus D^t consists of the elements $d_{\theta^*}{}^t$ corresponding to all real values θ^*. We shall put

$$W(\theta, d_{\theta^*}{}^t) = (\theta^* - \theta)^2 \tag{5.136}$$

The cost of experimentation is assumed to be proportional to the number of observations. Let c denote the cost of a single observation.

The problem considered here is again a special case of the general problem treated in Chapter 4. We shall deal here with the question of finding a minimax solution for our problem. For this purpose we shall first derive the Bayes solution when the a priori distribution of θ is a rectangular distribution with a given range $[a, b]$, where $a < b$. Suppose that m observations $x_1, \cdots, x_m$ have been made. Let

$$(5.137) \quad u = \text{Min}\,(x_1, \cdots, x_m) \qquad \text{and} \qquad v = \text{Max}\,(x_1, \cdots, x_m)$$

Then the a posteriori probability distribution of θ is again a rectangular distribution whose range is equal to the common part of the intervals $[a, b]$ and $[v - \frac{1}{2}, u + \frac{1}{2}]$. Let $I(u, v)$ denote the common part of $[a, b]$ and $[v - \frac{1}{2}, u + \frac{1}{2}]$, and let $t(u, v)$ denote the midpoint of the interval $I(u, v)$. It is clear that the optimum terminal decision is to estimate θ by $t(u', v')$, where u' and v' are the values of u and v, respectively, at the termination of the experimentation. Thus the problem of finding a Bayes solution reduces to the problem of finding an optimum rule for stopping experimentation.

Clearly, whether a rectangular distribution ξ in Ω is of type 1, type 2, or type 3 (for a definition of the three types see Section 4.1.3) depends only on the length l of the range of ξ. Thus it is possible to subdivide the non-negative half of the real axis into three disjoint sets R_1, R_2, and R_3 such that when l is a point in R_i the distribution ξ is of type i $(i = 1, 2, 3)$.

Let l_m denote the length of the interval $I(u, v)$ after m observations $x_1, \cdots, x_m$ have been made $(m = 1, 2, \cdots,$ ad inf.). We define l_0 as the length of the range of the a priori distribution; i.e., $l_0 = b - a$. With the help of the sets R_1, R_2, and R_3, a Bayes solution can be given as follows: At the mth stage of the experiment $(m = 0, 1, 2, \cdots,$ etc.), compute l_m. If l_m is a point in R_1, stop experimentation and make the proper terminal decision. If l_m is a point in R_2, we can decide at random between taking an additional observation and making the proper terminal decision. If l_m is a point of R_3, an additional observation is made. Thus the problem of constructing a Bayes solution reduces to the problem of determining the sets R_1, R_2, and R_3.

If experimentation is terminated with the mth observation, the a posteriori risk associated with the terminal decision is equal to the a posteriori expected value of $[\theta - t(u, v)]^2$ which is simply equal to $l_m^2/12$. For the purpose of determining the sets R_1, R_2, and R_3, it will be necessary to determine the conditional expected value of $l^2_{m+1}/12$ when $l_m = l$ and l is a given positive number $(m = 0, 1, 2, \cdots,$ ad inf.). A simple computation shows that the conditional

expected value in question is given by

$$E\left(\frac{l^2_{m+1}}{12}\,\middle|\, l_m = l\right) = \frac{l^2}{12} - \frac{l^3}{24} \tag{5.138}$$

when $O \leqq l \leqq 1$, and

$$E\left(\frac{l^2_{m+1}}{12}\,\middle|\, l_m = l\right) = \frac{1}{12} - \frac{1}{24l} \tag{5.139}$$

when $l \geqq 1$.

Let

$$\phi(l) = \frac{l^2}{12} - E\left(\frac{l^2_{m+1}}{12}\,\middle|\, l_m = l\right) \tag{5.140}$$

Then

$$\phi(l) = \frac{l^3}{24} \tag{5.141}$$

when $l \leqq 1$, and

$$\phi(l) = \frac{l^2 - 1}{12} + \frac{1}{24l} \tag{5.142}$$

when $l \geqq 1$. The quantity $\phi(l)$ is simply the expected decrease in the a posteriori risk associated with the terminal decision due to an additional observation when l is the length of the interval $I(u, v)$ before the additional observation is made. Clearly $\phi(l)$ is strictly increasing with increasing l over the whole range of l. Thus the equation in l

$$\phi(l) = c \quad (c = \text{cost of a single observation}) \tag{5.143}$$

has exactly one root. Let $l = \bar{l}$ be the root of this equation. Since $\phi(l)$ is a monotonic function of l, and since $l_{m+1} \leqq l_m$ ($m = 0, 1, 2, \cdots$, ad inf.), we can easily verify that R_1 consists of all values $l < \bar{l}$, R_2 contains the single value $\bar{l}$, and R_3 consists of all values $l > \bar{l}$.

Thus, if the a priori distribution in Ω is a rectangular distribution, a Bayes solution is given by the following rule: At the mth stage of the experiment, for each non-negative integral value m compute the value of l_m. If $l_m < \bar{l}$, stop experimentation and make the proper terminal decision. If $l_m > \bar{l}$, take an additional observation. If $l_m = \bar{l}$, we can decide at random between making a terminal decision and taking an additional observation.

Let δ_0 be the decision rule given as follows: At least one observation is made. Experimentation is stopped at the mth observation with the adoption of $(u_m + v_m)/2$ as the point estimate of θ, where m is the smallest positive integer for which $(u_m - v_m + 1) \leqq \bar{l}$, $u_m = \text{Min}\ (x_1,$

$\cdots, x_m)$, and $v_m = \text{Max}\,(x_1, \cdots, x_m)$. We shall show that δ_0 is a minimax solution of our decision problem. Clearly $r(\theta, \delta_0)$ is constant over the whole domain of θ. Let r_0 be the constant value of $r(\theta, \delta_0)$.

For any positive integer k, let ξ_k be the rectangular distribution in Ω with range $[-k, k]$. Let δ_k be the Bayes solution relative to ξ_k according to which experimentation is stopped as soon as $l_m \leqq \bar{l}$. For any $k > \text{Max}\,(2, \bar{l}/2)$, we have

$$r(\theta, \delta_k) = r(\theta, \delta_0) = r_0 \tag{5.144}$$

for $-(k-1) \leqq \theta \leqq k-1$. The above equation is an immediate consequence of the fact that δ_0 coincides with δ_k when $|\theta| \leqq k-1$. Suppose now that δ_0 is not a minimax solution. Then there exist a decision function δ^* and a positive value $r_0^* < r_0$ such that

$$r(\theta, \delta^*) \leqq r_0^* < r_0 \tag{5.145}$$

for all θ. Clearly

$$\limsup_{k=\infty} r(\xi_k, \delta^*) \leqq r_0^* \tag{5.146}$$

It follows from (5.144) that

$$\liminf_{k=\infty} r(\xi_k, \delta_k) \geqq r_0 \tag{5.147}$$

Since δ_k is a Bayes solution relative to ξ_k, equation (5.146) cannot hold. Thus we arrive at a contradiction, and our statement that δ_0 is a minimax solution is proved.

BIBLIOGRAPHY

1. Albert, G. E., "A Note on the Fundamental Identity of Sequential Analysis," *Ann. Math. Stat.*, **18** (1947).
1a. Albert, G. E., "Correction to 'A Note on the Fundamental Identity of Sequential Analysis,'" *Ann. Math. Stat.*, **19** (1948).
2. Anscombe, F. J., "Linear Sequential Rectifying Inspection for Controlling Fraction Defective," *Suppl. J. Roy. Stat. Soc.*, **8** (1946).
2a. Anscombe, F. J., Goodwin, H. J., and Plackett, R. L., "Methods of Deferred Sentencing in Testing the Fraction Defective of a Continuous Output," *Suppl. J. Roy. Stat. Soc.*, **9** (1947).
3. Armitage, P., "Sequential Tests of Student's Hypothesis," *Suppl. J. Roy. Stat. Soc.*, **9** (1947).
4. Arrow, K. J., D. Blackwell, and M. A. Girshick, "Bayes and Minimax Solutions of Sequential Decision Problems," *Econometrica*, **17** (1949).
5. Banach, S., "Théorie des opérations linéaire," *Monografje Matematyczne*, Warszawa, I (1932).
6. Barnard, G. A., "Sequential Tests in Industrial Statistics," *Suppl. J. Roy. Stat. Soc.*, **8** (1946).
7. Bartky, W., "Multiple Sampling with Constant Probability," *Ann. Math. Stat.*, **14** (1943).
8. Bartlett, M. S., "The Large-Sample Theory of Sequential Tests," *Proc. Cambridge Phil. Soc.*, **42** (1946).
9. Blackwell, D., and M. A. Girshick, "On Functions of Sequences of Independent Chance Vectors with Applications to the Problem of 'Random Walk' in k Dimensions," *Ann. Math. Stat.*, **17** (1946).
10. Blackwell, David, "On an Equation of Wald," *Ann. Math. Stat.*, **17** (1946).
11. Blackwell, David, "Conditional Expectation and Unbiased Sequential Estimation," *Ann. Math. Stat.*, **18** (1947).
12. Blackwell, D., and M. A. Girshick, "A Lower Bound for the Variance of Some Unbiased Sequential Estimates," *Ann. Math. Stat.*, **18** (1947).
13. Burman, J. P., "Sequential Sampling Formulae for a Binomial Population," *Suppl. J. Roy. Stat. Soc.*, **8** (1946).
14. Dodge, H. F., and H. G. Romig, "A Method of Sampling Inspection," *Bell System Tech. J.*, **8** (1929).
15. Fisher, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans. Roy. Soc.*, **A222** (1921).
16. Fisher, R. A., "Theory of Statistical Estimation," *Proc. Cambridge Phil. Soc.*, **22** (1925).
17. Fisher, R. A., "The Logic of Inductive Inference," *J. Roy. Stat. Soc.*, **98** (1935).
18. Fisher, R. A., *The Design of Experiments*, Oliver and Boyd, London, 3rd ed., 1942.
19. Girshick, M. A., "Contributions to the Theory of Sequential Analysis: I," *Ann. Math. Stat.*, **17** (1946).
20. Girshick, M. A., "Contributions to the Theory of Sequential Analysis: II, III," *Ann. Math. Stat.*, **17** (1946).

21. Girshick, M. A., F. Mosteller, and L. J. Savage, "Unbiased Estimates for Certain Binomial Sampling Problems with Applications," *Ann. Math. Stat.*, **17** (1946).
22. Harris, T. E., "Note on Differentiation under the Expectation Sign in the Fundamental Identity of Sequential Analysis," *Ann. Math. Stat.*, **18** (1947).
23. Hausdorff, F., *Mengenlehre*, Walter de Gruyter & Co., Berlin, 1927.
24. Helly, E., "Über Systeme linearer Gleichungen mit unendlich vielen Unbekannten," *Monatshefte Math. Physik*, **31** (1921).
25. Herbach, L. H., "Bounds for Some Functions Used in Sequentially Testing the Mean of a Poisson Distribution," *Ann. Math. Stat.*, **19** (1948).
26. Kac, Mark, "Random Walk and the Theory of Brownian Motion," *Amer. Math. Monthly*, **54** (1947).
27. Kaplansky, I., "A Contribution to von Neumann's Theory of Games," *Ann. Math.*, **46** (1945).
28. Kryloff, N., and N. Bogoliouboff, "La théorie générale de la mesure dans son application à l'étude des systèmes dynamiques de la mécanique non-linéaire," *Ann. Math.*, **38** (1937).
29. Lefschetz, Solomon, "Algebraic Topology," *Amer. Math. Soc. Colloquium Publ.*, **27** (1942).
30. Lehmann, E. L., "On Families of Admissible Tests," *Ann. Math. Stat.*, **18** (1947).
31. Mahalanobis, P. C., "A Sample Survey of the Acreage under Jute in Bengal with Discussion of Planning of Experiments," *Proc. 2nd Indian Stat. Conf.*, Calcutta, Statistical Publishing Society, 1940.
32. Milgram, A. N., "Partially Ordered Sets, Separating Systems, and Inductiveness," *Reports of Mathematical Colloquium*, 2nd Series, Issue 1, University of Notre Dame Press, South Bend, Ind., 1939.
33. Nandi, H. K., "Use of Well Known Statistics in Sequential Analysis," *Sankhyā*, **8**, Pt. 4 (June 1948).
34. Neyman, J., and E. S. Pearson, "The Testing of Statistical Hypotheses in Relation to Probability a Priori," *Proc. Camb. Phil. Soc.*, **29** (1933).
35. Neyman, J., "Sur la vérification des hypothèses statistiques composées," *Bull. soc. math. France*, **63** (1935).
36. Neyman, J., "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Phil. Trans. Roy. Soc.*, **A236** (1937).
37. Neyman, J., and E. S. Pearson, "Contributions to the Theory of Testing Statistical Hypotheses," *Stat. Res. Mem.*, Pts. I and II (1936, 1938).
38. Neyman, J., "L'estimation statistique traitée comme un problème classique de probabilité," *Actualités sci. ind.*, No. 739 (1938).
39. Paulson, Edward, "A Note on the Efficiency of the Wald Sequential Test," *Ann. Math. Stat.*, **18** (1947).
40. Pitman, E. J. G., "The 'Closest' Estimate of Statistical Parameters," *Proc. Cambridge Phil. Soc.*, **33** (1937).
41. Pitman, E. J. G., "The Estimation of Location and Scale Parameters of a Continuous Population of any Given Form," *Biometrika*, **30** (1939).
42. Plackett, R. L., "Boundaries of Minimum Size in Binomial Sampling," *Ann. Math. Stat.*, **19** (1948).
42a. Pólya, George, *Exact Formulas in the Sequential Analysis of Attributes*, University of California Press, 1948.
43. Robbins, Herbert, "Convergence of Distributions," *Ann. Math. Stat.*, **19** (1948).

43a. Robinson, Julia, *A Note on Exact Sequential Analysis*, University of California Press, 1948.
44. Saks, S., *Theory of the Integral*, Hafner Publishing Company, New York, 1937.
45. Samuelson, Paul A., "Exact Distribution of Continuous Variables in Sequential Analysis," *Econometrica*, **16** (1948).
46. Savage, L. J., "A Uniqueness Theorem for Unbiased Sequential Binomial Estimation," *Ann. Math. Stat.*, **18** (1947).
47. Silber, J., "Multiple Sampling for Variables," *Ann. Math. Stat.*, **19** (1948).
48. Sobel, M., "Complete Classes of Decision Functions for Some Standard Sequential and Non-Sequential Problems" (to be published).
48a. Sobel, M., and Wald, A., "A Sequential Decision Procedure for Choosing One of Three Hypotheses Concerning the Unknown Mean of a Normal Distribution," *Ann. Math. Stat.*, **20** (1949).
48b. Statistical Research Group, Columbia University, *Sequential Analysis of Statistical Data: Applications*, Columbia University Press, New York, 1945.
49. Stein, Charles, "A Two-Sample Test for a Linear Hypothesis whose Power Is Independent of the Variance," *Ann. Math. Stat.*, **17** (1945).
50. Stein, Charles, "A Note on Cumulative Sums," *Ann. Math. Stat.*, **17** (1946).
51. Stein, Charles, and A. Wald, "Sequential Confidence Intervals for the Mean of a Normal Distribution with Known Variance," *Ann. Math. Stat.*, **18** (1947).
52. Stein, C. M., "On Sequences of Experiments," abstracted in *Ann. Math. Stat.*, **19** (1948).
53. Stockman, C. M., and P. Armitage, "Some Properties of Closed Sequential Schemes," *Suppl. J. Roy. Stat. Soc.*, **8** (1946).
54. Ville, J., "Note 'Sur la théorie générale des jeux on intervient l'habilité des jouers,' " in the book *Applications aux jeux de hasard*, by Emile Borel and Jean Ville, Tome IV, Fascicule II of the *Traité du calcul des probabilités et de ses applications*, par Emile Borel (1938).
55. von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behaviour*, Princeton University Press, Princeton, 1944.
56. Wald, A., "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *Ann Math. Stat.*, **10** (1939).
57. Wald, A., "On Cumulative Sums of Random Variables," *Ann. Math. Stat.*, **15** (1944).
58. Wald, A., "Generalization of a Theorem by von Neumann Concerning Zero Sum Two Person Games," *Ann. Math.*, **46** (1945).
59. Wald, A., "Statistical Decision Functions which Minimize the Maximum Risk," *Ann. Math.*, **46** (1945).
60. Wald, A., "Some Generalizations of the Theory of Cumulative Sums of Random Variables," *Ann. Math. Stat.*, **16** (1945).
61. Wald, A., "Sequential Tests of Statistical Hypotheses," *Ann. Math. Stat.*, **16** (1945).
61a. Wald, A., "Sequential Method of Sampling for Deciding Between Two Courses of Action," *J. Amer. Stat. Assoc.*, **40** (1945).
62. Wald, A., "Some Improvements in Setting Limits for the Expected Number of Observations Required by a Sequential Probability Ratio Test," *Ann. Math. Stat.*, **17** (1946).
63. Wald, A., "Differentiation under the Expectation Sign in the Fundamental Identity of Sequential Analysis," *Ann. Math. Stat.*, **17** (1946).
64. Wald, A., "Limit Distribution of the Maximum and Minimum of Successive Cumulative Sums of Random Variables," *Bull. Amer. Math. Soc.*, **53** (1947).

65. Wald, A., *Sequential Analysis*, John Wiley & Sons, Inc., New York, 1947.
66. Wald, A., "An Essentially Complete Class of Admissible Decision Functions," *Ann. Math. Stat.*, **18** (1947).
67. Wald, A., "Foundations of a General Theory of Sequential Decision Functions," *Econometrica*, **15** (1947).
68. Wald, A., "On the Distribution of the Maximum of Successive Cumulative Sums of Independently but not Identically Distributed Chance Variables," *Bull. Amer. Math. Soc.*, **54** (1948).
69. Wald, A., and J. Wolfowitz, "Optimum Character of the Sequential Probability Ratio Test," *Ann. Math. Stat.*, **19** (1948).
70. Wald, A., "Statistical Decision Functions," *Ann. Math. Stat.*, **20** (1949).
71. Wald, A., and J. Wolfowitz, "Bayes Solution of Sequential Decision Problems," *Ann. Math. Stat.*, **21** (1950).
72. Widder, David Vernon, *The Laplace Transform*, Princeton University Press, Princeton, 1946.
73. Wolfowitz, J., "On Sequential Binomial Estimation, "*Ann. Math. Stat.*, **17** (1946).
74. Wolfowitz, J., "Consistency of Sequential Binomial Estimates," *Ann. Math. Stat.*, **18** (1947).
75. Wolfowitz, J., "The Efficiency of Sequential Estimates and Wald's Equation for Sequential Processes," *Ann. Math. Stat.*, **18** (1947).
75a. Wolfowitz, J., "Minimax Estimates of the Mean of a Normal Distribution with Known Variance," *Ann. Math. Stat.*, **21**, No. 2 (1950).
76. Zorn, Max, "A Remark on Method in Transfinite Algebra," *Bull. Amer. Math. Soc.*, **41** (1935).

INDEX