12

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 • U.S.A.

THE INVARIANCE APPROACH TO THE PROBABILISTIC ENCODING OF INFORMATION

by

Daniel Warner North

Decision Analysis Group

Stanford Research Institute

Menlo Park, California

March 1970

ACKNOWLEDGMENTS

A dissertation is an individual effort, and yet many debts to others are incurred in the process. Perhaps my greatest debt is to my friends, and especially to my parents. Without their encouragement and support I would never have made it.

I am indebted to Professors George Murray and Robert Wilson for asking the questions that led to this research as well as for serving on my dissertation reading committee. In addition, my colleagues from the Decision Analysis Group at Stanford Research Institute provided much helpful criticism of the dissertation manuscript. Mrs. Sheila Hill did an exemplary job on the typing. Of course, the responsibility for the final contents of the dissertation, including any errors of commission or omission, is entirely mine. I am grateful to the Department of Operations Research and to the National Science Foundation for fellowship support during the early stages of the research.

To my adviser, Professor Ronald Howard, I owe a special debt of thanks. It has been a privilege to do my dissertation under his guidance.

ABSTRACT

The idea that probability assignments reflect a state of information is fundamental to the use of probability theory as a means of reasoning about uncertain events. In order to achieve a consistent methodology for assigning probabilities the following basic desideratum is required: Two states of information that are perceived to be equivalent should lead to the same probability assignments. This basic desideratum leads to an invariance approach to the probabilistic encoding of information, because the probability assignments must remain invariant to a change from one state of information to an equivalent state of information. The criterion of insufficient reason is one application of the invariance approach.

The principle of maximum entropy has been proposed as a general means of assigning probabilities on the basis of specified information. Invariance considerations provide a much stronger justification for this principle than has been heretofore available. Statistical equilibrium (invariance to randomization over time) provides the basis for the maximum entropy principle in statistical mechanics. This derivation allows a complete and comprehensive exposition of J. Willard Gibbs' approach to statistical mechanics to be formulated for the first time.

Repeated, indistinguishable experiments have been the traditional concern of statistics. De Finetti's concept of exchangeability is an invariance principle that connects the inductive use of probability to the traditional relative frequency viewpoint. An extension of the criterion of insufficient reason to exchangeable sequences provides a basis for the maximum entropy principle. The derivation provides new insight into the process of inference using sufficient statistics. The Koopman-Pitman theorem relates distributions characterized by sufficient statistics to states of information concerned with long run averages.

The invariance approach gives other insights into the use of probability theory as well. Exchangeability can be applied to time-dependent and conditional random processes; infinitely divisible processes are an interesting special case. Since the invariance approach is based on the perceived equivalence between states of information, it is important to have a means for questioning an assumed equivalence as further information becomes available. A method for questioning and revising equivalence assumptions is given, and its relation to the classical theory of statistical hypothesis testing is discussed.

TABLE OF CONTENTS

Chapter I

INTRODUCTION

1.1 The Epistemological Controversy

One of the most enduring of all philosophical controversies has concerned the epistemology of uncertainty: how can logical methods be used to reason about what is not known? When a formal theory of probability was developed between two and three hundred years ago, it was hailed as the answer to this question. Laplace wrote in the introductionto A Philosophical Essay on Probabilities ([47], p. 1):

> I have recently published upon the same subject a work entitled The Analytical Theory of Probabilities. I present here without the aid of analysis the principles and general results of this theory, applying them to the most important questions of life, which are indeed for the most part only problems of probability. Strictly speaking it may even be said that nearly all our knowledge is problematical; and in the small number of things which we are able to know with certainty, even in the mathematical sciences themselves, the principal means for ascertaining truth - induction and analogy - are based on probabilities; so that the entire system of human knowledge is connected with the theory set forth in this essay.

This viewpoint did not prevail for long. Later writers restricted the domain of probability theory to repetitive situations analogous to games of chance. The probability assigned to an event was defined to be the limiting fraction of the number of times the event occurred in a large number of independent, repeated trials. This "classical" viewpoint underlies most of statistics, and it is still widely held among contemporary probabilists.

In recent years the broader view held by Laplace has reemerged. Two different approaches have led to the resurgence of probability theory as a general way of reasoning about uncertainty. Ramsey [72] developed a personalistic theory of probability as a means of guaranteeing consistency for an individual's choices among wagers. Savage [74] combined Ramsey's ideas with the von Neumann and Morgenstern [90] theory of risk preference to achieve a formulation of decision theory in which the axioms of probability emerge as the basis for representing an individual's degree of belief in uncertain events.

Savage's formulation is behavioralistic because it relies on the individual's choice among wagers as the operational means for measuring probability assignments. A subject is asked to choose between wagering on an uncertain event and wagering on a probabilistic "reference" process for which the odds of winning are clearly evident, for example, from symmetry considerations.* If the odds for the reference process are adjusted so that the subject believes that the two wagers are of equal value, then this number may be taken as a summary of his subjective judgment about the occurrence of the uncertain event. Probabilities determined in this fashion are often called subjective or personal probabilities.

The other approach adopts a logical rather than a decision-oriented viewpoint. We desire a means for reasoning logically about uncertain events; we are not concerned with making decisions among wagers. The axioms of probability emerge as a consequence of intuitive assumptions

* For a more detailed discussion of encoding probability assignments, see North [60].

about inductive reasoning. This viewpoint was held by Keynes [44] and Jeffreys [43]. Perhaps the most convincing development is the functional equation derivation of R. T. Cox: If one is to extend Boolean algebra so that the uncertainty of events may be measured by real numbers, consistency requires that these numbers satisfy conditions equivalent to the axioms of probability theory ([9], [35], [85]).

An essential feature of both the personalistic and the "logical" approaches is that probabilities must reflect the information upon which they are based. Probability theory gives a way of reasoning about one state of information in terms of other states of information. The means by which this reasoning is accomplished is Bayes' Rule, a simple formula that is equivalent to the multiplication law for conditional probabilities. Since Bayes' Rule plays such an important role, probability theory as a general means of reasoning about uncertainty is often distinguished by the adjective "Bayesian" from the more limited relative frequency or "classical" use of probability theory.

The major difference between the personalistic and the "logical" approach is in the starting point: the prior probabilities. How are probabilities to be initially assigned? The personalist has a ready answer: ask the subject to make a choice among wagers. The "logical" school replies by asking why we should assume that the subject will make choices that reflect his true state of knowledge. Substantial evidence exists that people do not process information in a way consistent with the laws of probability (for example, Edwards et al [13], Raiffa [69]). A logical means of reasoning about uncertainty should be free of irrational or capricious subjective elements, and there is no assurance that the personalistic theory fulfills this requirement. The personalist counters that there is no alternative to subjective judgment available for assigning probabilities. It is simply not possible to start by assuming no knowledge at all and then process all information by Bayes' Rule (Jeffreys [43], p. 33). Unless there is some formal method for translating information into probability assignments we shall be forced to rely on the subjective judgment of the individual.*

1.2 The Basic Desideratum

In this dissertation we shall examine a method for translating information into probability assignments. For our efforts to have meaning we shall require the following assumption, which seems self-evident for both the personalistic and "logical" approaches to Bayesian probability:

> A probability assignment reflects a state of information.

In the personalistic approach the process of translating information into probability assignments is left to the individual. For the "logical" approach it would be highly desirable to have formal principles by which information might be translated into probability assignments. Using these principles would prevent arbitrary subjective elements from being introduced in going from the information to the probability assignment.

* Further exposition on the controversy surrounding prior information is to be found in such sources as Jeffreys [43], Savage [74], [75], and Jaynes [40].

A basis for developing such formal principles has been suggested by Jaynes [40]. As a basic desideratum let us require that "in two problems where we have the same prior information, we should assign the same prior probabilities." The intended implication of this statement is that a state of knowledge should lead to a unique choice of models and probability distributions, independent of the identity, personality, or economic status of the decision-maker. The essential element of the basic desideratum is the notion of invariance. If two or more states of information are judged equivalent, the probability distribution should not depend on which of the states of information was chosen as the basis for encoding the probability distribution. The probability distribution should remain invariant if one of these states of information is replaced by another.

This invariance approach to the probabilistic encoding of information does not resolve all the difficulties. We cannot avoid subjective elements in the process of encoding a probability distribution to represent a state of information. The notion of equivalence between states of information is always an approximation. No two distinct situations can be the same in all aspects. Where more than one individual is concerned we must consider the difference in background. Two people will judge a given situation on the basis of information and prior experiences that are necessarily different in some particulars. Before we can use the basic desideratum we must decide what information and experience is relevant in making a probability assignment.

No criterion by which irrelevant information might be discarded appears to be available, save subjective judgment based on past experience. The basic desideratum does not provide the basis for a truly "objective" methodology, for it must rely on an individual's subjective judgment that two states of information are equivalent. This equivalence may be clearly intuitive in some situations, while in others it represents only a crude approximation to the individual's judgment. Furthermore, an equivalence that was initially assumed may appear doubtful after further information has become available that distinguishes one state of information from another.

The necessity to use subjective judgment in applying the basic desideratum suggests that we rephrase it as follows:

> The Basic Desideratum: Two states of information that are perceived to be equivalent should lead to the same probability assignments.

We cannot eliminate subjective elements completely, but by using this basic desideratum we move these elements back one stage. Instead of in the assignment of probabilities, subjective elements appear in the way that we characterize the relevant information. It is doubtful that the quest for an objective methodology can be pushed much further.

1.3 An Overview of the Invariance Approach

The basic desideratum as we have stated it above is so simple that it appears to be a tautology. We shall see that it provides a unifying basis for principles to translate information into probability distributions and probabilistic models. Many of these principles have been available for years in the writings of Laplace [47], Gibbs [23], and more recently, de Finetti [20], [21]. Perhaps the most important contribution is that of Jaynes [33], [34], [35], [36], [37], [38], [39], [40], [41]: the maximum entropy principle for assigning probabilities on the basis of explicitly stated prior information.

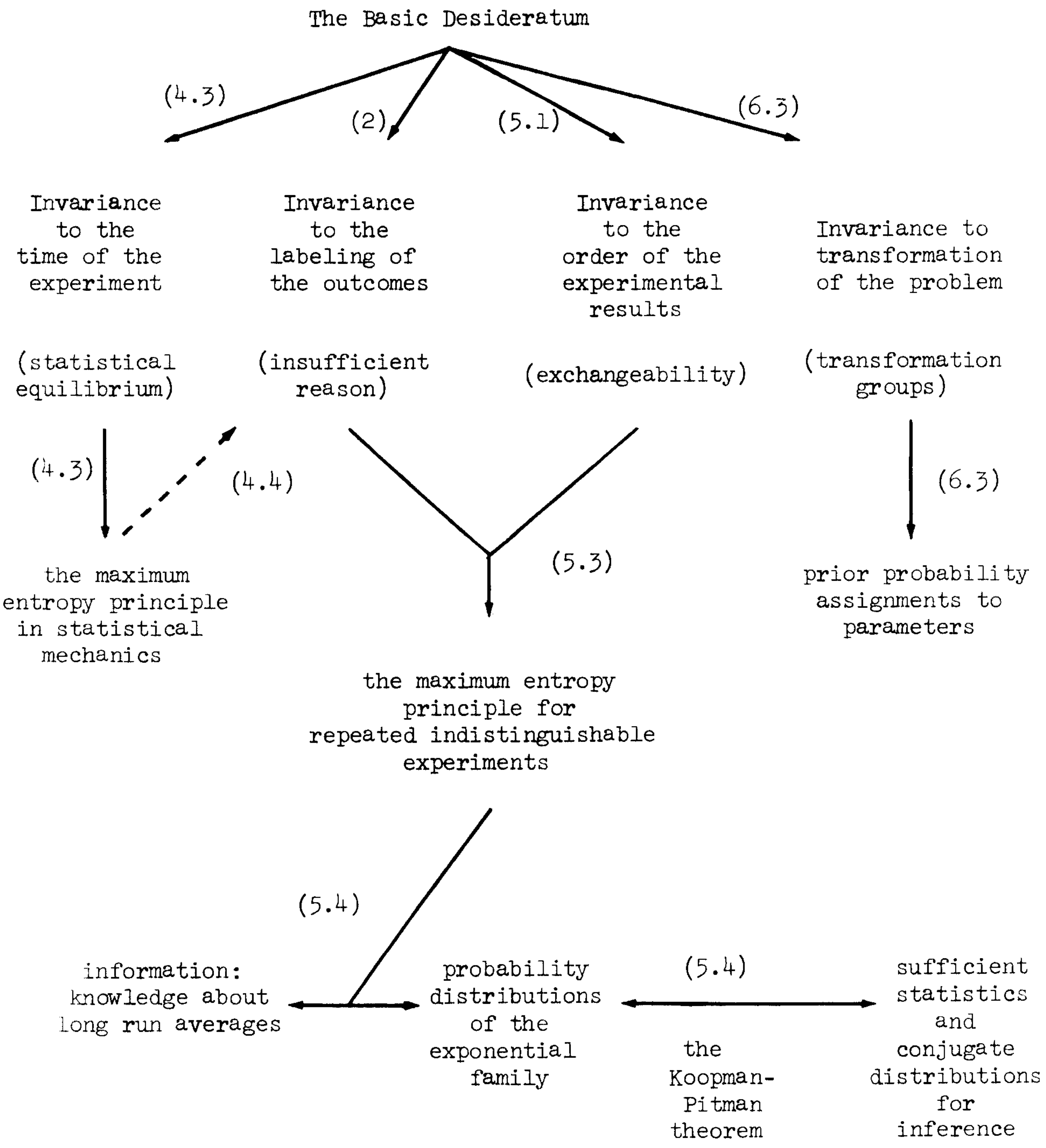
This dissertation has been largely stimulated by Jaynes' work and in many respects it is an extension of the lines of investigation that he has pioneered. There is one important difference in emphasis. Whereas Jaynes ascribes a fundamental importance to entropy, the present work is based on invariance criteria derived from the basic desideratum stated above. From these invariance criteria Jaynes' maximum entropy principle may be derived.

We shall examine four applications of the basic desideratum. An overview of these applications is given in Figure 1.1; the numbers in parenthesis refer to the sections where the corresponding connection is discussed.

Chapter 2 is devoted to the criterion of insufficient reason, which is obtained by assuming that the state of information remains invariant under a relabeling of the possible outcomes in an uncertain situation. The discussion is illustrated using the Ellsberg paradox as an example.

A review of the axiomatic approach to the maximum entropy principle in Chapter 3 shows a need to derive this principle from more fundamental considerations. Such a derivation is accomplished in Chapters 4 and 5 for two quite different problems using different applications of the basic desideratum. Chapter 4 is devoted to the problem in statistical mechanics addressed by J. Willard Gibbs. In Gibbs' deterministic problem the uncertainty concerns the initial conditions for a set of differential equations of motion. The invariance is to the time at

Figure 1.1: Overview of the Invariance Approach


The Basic Desideratum
(4.3)
(2)
(5.1)
(6.3)
Invariance to the time of the experiment
Invariance to the labeling of the outcomes
Invariance to the order of the experimental results
Invariance to transformation of the problem
(statistical equilibrium)
(insufficient reason)
(exchangeability)
(transformation groups)
(4.3)
(4.4)
(5.3)
(6.3)
the maximum entropy principle in statistical mechanics
prior probability assignments to parameters
the maximum entropy principle for repeated indistinguishable experiments
(5.4)
(5.4)
information: knowledge about long run averages
probability distributions of the exponential family
sufficient statistics and conjugate distributions for inference
the Koopman-Pitman theorem

which these initial conditions are determined; the concept of statistical equilibrium implies that the time of the determination should not affect our information about the physical system. In particular, our information should be the same whether the initial conditions are determined at a fixed time or at a time chosen through some random mechanism. This invariance to randomization provides a direct proof for the principle of maximum entropy in statistical mechanics. The principle has a relation to the criterion of insufficient reason that we shall examine in the last section of Chapter 4.

Chapter 5 begins the exploration of another application of the basic desideratum, de Finetti's concept of exchangeability. This concept provides a Bayesian interpretation to the notion of a statistical ensemble. The invariance is to changes in the order in a sequence of experimental results. De Finetti's theorem provides an important insight into the nature of inference on repeated, indistinguishable experiments. If the number of possible experimental outcomes is large an additional principle is needed to reduce the inference problem to manageable proportions. One such principle is the criterion of insufficient reason extended to sequences of experimental outcomes; from this the principle of maximum entropy may be derived.

In the special case in which the state of knowledge concerns only long run averages the maximum entropy principle for exchangeable sequences leads to a special form for the probability distribution: the exponential family. The Koopman-Pitman theorem shows that membership in the exponential family is a necessary and sufficient condition for inference using sufficient statistics. The case in which available knowledge concerns fractiles of the probability distribution and possibly long run averages as well is solved using the maximum entropy principle and standard optimization methods.

Chapter 6 examines the significance of the relationship between sufficient statistics, the exponential family, and information in the form of averages. By examining this relationship in the light of the extended version of the criterion of insufficient reason, we gain considerable insight into the conceptual basis for probability distributions having sufficient statistics and therefore permitting conjugate distributions for inference on parameters.

The basic desideratum can be applied directly to the determination of prior probability distributions on parameters. If the problem of assigning a distribution to a set of parameters is perceived to be equivalent to the problem of assigning a distribution to a second set of parameters related to the first set by a functional transformation, a functional equation may be developed to solve for the probability distribution on the parameter set. This method of transformation groups is the only application of the basic desideratum discussed by Jaynes [41]; it is summarized in Section 6.3.

The final application of the invariance approach, in Section 6.4, is to more complex repetitive processes. Exchangeability applied to a continuous process leads to infinite divisibility, and a very strong characterization of the process results. Exchangeability may be weakened to conditional exchangeability if only some reorderings of experimental results result in new states of information that are perceived to be equivalent to the original state of information.

Conditional exchangeability provides a basis for time-dependent and Markov dependent probabilistic models.

The basic desideratum can be applied only when states of information are perceived to be equivalent; further information about an uncertain situation may throw the equivalence into doubt. In Chapter 7 a means is developed for testing whether probabilistic models derived from the basic desideratum are still appropriate after further information has been received. The test is based on the use of Bayes' Rule, but it can be related to some of the classical methods for statistical hypothesis testing.

Chapter II

THE CRITERION OF INSUFFICIENT REASON

Let us suppose that we are able to specify the information that is or is not relevant to the outcome of an uncertain situation. We would like a principle detailing how a state of information should specify a probability distribution. One such principle is the criterion of insufficient reason, formulated by Jacob Bernoulli in the seventeenth century and adopted later by Bayes and Laplace.

The criterion of insufficient reason may be viewed as an invariance principle that characterizes certain states of information. We might describe these states of information as "maximum ignorance" and we shall define this property as follows: given an outcome space* of N (mutually exclusive, collectively exhaustive) possible outcomes, our state of information remains the same if two or more of the outcomes are interchanged; we perceive the problems as being identical before and after the relabeling of the outcomes. The criterion of insufficient reason states that for such a state of information, equal probabilities should be assigned to each outcome. We can see that this criterion follows immediately from the basic desideratum that probability assignments are determined by the state of information. If our information remains invariant under an interchange of elements in the outcome space,

* We prefer this term to the "set of states of the world" used by Savage [74] and others, and we wish to avoid the frequency connotation and possible confusion with results of information-gathering inherent in "sample space" or "sample description space" used by Feller [17], Parzen [61] et al.

the probability assignments must remain unchanged by such a relabeling. Hence, the probability assigned to any outcome must equal the probability assigned to any other outcome; we must assign all outcomes equal probability.*

To call a state of information that possesses the invariance property described above "maximum ignorance" is somewhat misleading, for we have encoded a considerable sum of knowledge in choosing the set of outcomes that constitute the outcome space. Most of the confusion and criticism of the criterion of insufficient reason has resulted from failure to recognize the essential first step in the use of probability theory: specification of an outcome space, a "universe of discourse" that constitutes the set of possible outcomes or events to which probabilities will be assigned. Vague terms such as "states of the world" and "states of nature" tend to obscure this fundamental aspect of the encoding process. Before someone can meaningfully assign a probability to an event it must be clear to him exactly what the event is, and what set of events constitutes its complement.

To understand the applicability of the criterion of insufficient reason, we must gain some feeling for what constitutes "maximum ignorance." A good starting point might be to examine some examples proposed by Ellsberg [15].

Consider the problem of assigning a probability p to drawing a ball of the color we have chosen from an urn filled with black balls

* An invariance interpretation of the criterion of insufficient reason has been discussed in a slightly different context by Chernoff [6].

and red balls. In one case (problem I) the urn is known to contain one hundred balls, but nothing is known about the proportion of black and red. In a second case (problem II) the urn is known to contain exactly fifty black balls and fifty red balls. For both problems it is clear that the outcome space is composed of two events: R, the event that a red ball is drawn, and B, the event that a black ball is drawn. If we choose the color "red," then we have picked $p = p(R)$ as the probability to be assigned, and from the axioms of probability we deduce $p(B) = 1 - p$. But suppose we choose the color "black;" both problem I or problem II remain exactly the same except now $p = p(B)$, and $p(R) = 1 - p$. By choosing black instead of red, we are in effect relabeling the states (Figure 2.1):

	Choose "red"	Choose "black"
chosen color is drawn	red	black
other color is drawn	black	red

Figure 2.1: Relabeling the State Space

We feel that the result of the draw does not depend on which color we chose, and so our information is invariant with respect to the change in the labeling of the outcomes. Then for both problem I and problem II we may characterize our state of information as "maximally ignorant." By the criterion of insufficient reason we should assign a probability $p = 1/2$ to drawing a red ball if we choose red, or to a black ball if we choose black; i.e., regardless of our choice of color $p(R) = p(B) = 1/2$. We could state in the same way that our state of information is one of "maximum ignorance" with regard to the outcome of heads versus tails on the flip of a coin, if interchanging the labels "heads" and "tails" for the two sides of the coin results in a new state of information that we judge to be indistinguishable from the original state of information. The generalization of the concept to more than two states is straightforward, if one keeps in mind that a "maximum ignorance" state of information must be invariant to any permutation of the states in the outcome space.

Ellsberg was concerned with the empirical phenomenon that people prefer to place bets on problem II rather than problem I; they seem to exhibit a preference for "known" probabilities as opposed to "unknown" probabilities. The volume of controversy that his paper has stimulated ([15], [18], [71], [5], [73], [19], [80]) is perhaps indicative of the confusion that still pervades the probabilistic foundations of decision theory. Part of the confusion stems from indecision as to whether decision theory should be normative or descriptive. If a normative posture is adopted, then an argument advanced by Raiffa [71] should convince us that for problem I as well as problem II we can choose the color so that the probability of drawing the color we choose is 1/2: Flip a coin and let the outcome determine the choice of "black" or "red." From the (subjective) assumption that the coin is "fair" we conclude that the probability we shall choose correctly the color of the ball to be drawn is 1/2. This randomization argument lends an intuitive meaning to the concept of a "maximum ignorance" state of information: Surely we can never be more ignorant than in the situation where the labels on the states in the outcome space are placed according to some "random"(i.e., equally probable outcomes) mechanism.* However, the same distribution of equal probability for each state may also characterize situations in which we feel intuitively that we have a great deal of knowledge. For example, in problem II we know the exact proportion of black and red balls in the urn, yet

* The existence of at least one such mechanism is an assumption that virtually everyone accepts, regardless of their views on the foundations and applicability of probability theory. Pratt, Raiffa, and Schlaifer ([65], p. 355) take this assumption (in the mind of the decision maker) as the basis for their development of subjective probability. De Finetti has an extremely cogent discussion on this point, which merits quoting in full ([20], p. 112):

> Thus in the case of games of chance, in which the calculus of probability originated, there is no difficulty in understanding or finding very natural the fact that people are generally agreed in assigning equal probability to the various possible cases, through more or less precise, but without doubt, very spontaneous, considerations of symmetry. Thus the classical definition of probability, based on the relation of the number of favorable cases to the number of possible cases, can be justified immediately: indeed, if there is a complete class of n incompatible events, and if they are judged equally probable, then by virtue of the theorem of total probability each of them will necessarily have the probability $p = 1/n$ and the sum of m of them the probability m/n. A powerful and convenient criterion is thus obtained: not only because it gives us a way of calculating the probability easily when a subdivision into cases that are judged equally probable is found, but also because it furnishes a general method for evaluating by comparison any probability whatever, by basing the quantitative evaluation on purely qualitative judgments (equality or inequality of two probabilities). However this criterion is only applicable on the hypothesis that the individual who evaluates the probabilities judges the cases considered equally probable; this is again due to a subjective judgment for which the habitual considerations of symmetry which we have recalled can furnish psychological reasons, but which cannot be transformed by them into anything objective.

we assign the same probabilistic structure as we do in problem I where we know "nothing" about the proportion.

The Ellsberg paradox points out very clearly that there are two dimensions to uncertainty which must be kept separate if we are to avoid confusion. Dimension one involves the assignment of probabilities to uncertain outcomes or states, while the second dimension measures a strength of belief in these assignments: how much the assignments would be revised as a result of obtaining additional information. Problems I and II are equivalent in dimension one; both have the equal probability assignments to outcomes that corresponds to the state of information, "maximum ignorance." However, the problems are vastly different in dimension two. Observing a red ball drawn (with replacement) from the urn in problem I will change the state of information into one for which the criterion of insufficient reason no longer applies. In problem II even the observation of many red balls drawn successively (with replacement) will not change the assignment of equal probability to drawing a red and a black ball on the next draw.*

The second dimension of uncertainty, how probability assignments will change with new information, lends itself readily to analytical treatment. Bayes' Rule provides a logical and rigorous framework for the revision of probabilities as additional information becomes available. Bayes' Rule requires, however, a conditional probability structure relating this additional information to the original state. In other words, we must have a likelihood function in order to use

* Of course, there comes a point at which we might question the model, e.g., the implicit assumption that any ball in the urn is equally likely to be selected. See Chapter 7.

Bayes' Rule. This likelihood function is nothing but another probability assignment, to which invariance principles such as the criterion of insufficient reason may or may not apply. Once we have the probabilistic structure needed for Bayes' Rule, we simply work through the mathematics of probability theory.

The criterion of insufficient reason and the generalizations that we shall discuss are in no way contradictory to Bayes' Rule and the axioms of probability; they serve as a means of determining which probabilistic structure will be an appropriate representation of the uncertainty in a given situation. We stress the essential prerequisite: the outcome space (the final outcomes as well as the possible results of further information-gathering) must be specified in advance.

Chapter III

THE MAXIMUM ENTROPY PRINCIPLE

Before we embark on the development of invariance principles as a basis for probability assignments it is advisable to motivate this development by examining some of the weaknesses in existing theory. Jaynes' maximum entropy principle provides a means of translating information into probability assignments, but, as we shall see, this approach has several difficulties.

Derivations of entropy as a measure of the information contained in a probability distribution have relied on the assumption that the information measure should be additive where more than one uncertain situation is considered. We present a slightly different derivation that proceeds from the assumption that the information measure should have the form of an expectation over possible outcomes. This approach is then compared to the traditional derivations that take additivity of the information measure as the fundamental assumption. The lack of intuitive justification for either the expectation or the additivity assumption implies the maximum entropy principle has not yet been given a secure conceptual foundation.

If we assume that entropy is the proper measure of the information contained in a probability distribution, then we can use entropy to choose which of several probability distributions should be assigned to represent a given state of information. Jaynes' maximum entropy principle is to choose the distribution consistent with given information whose entropy is largest. We examine the maximum entropy principle as it relates to methods of encoding probability distributions, and we find that it does not lead to interesting results except in the special case where some of the information concerns long run averages.

Later, in Chapters 5 and 6, we shall provide a secure conceptual foundation for the maximum entropy principle by deriving it from invariance principles, and we shall examine in detail the significance of long run average information.

3.1 Derivation of Entropy as a Criterion for Choosing Among Probability Distributions

Our task is to determine a measure of the information contained in a probability distribution. We noted in the last chapter that invariance with respect to relabeling of the states in the outcome space leads to an assignment of equal probability for each state, and intuitively we feel that this assignment represents a condition suitable to call "maximum ignorance," for we can always achieve this state of information by randomizing: selecting the i^{th} label for the j^{th} state with probability $1/N$, $j = 1, \ldots , N$. We now wonder if it is possible to develop a general measure of the "ignorance" inherent in a probability distribution. Shannon [78] first showed that such a measure of ignorance is specified by a few simple assumptions.

Consider an uncertain situation with N possible outcomes $A_1, \ldots , A_N$. This is the basic probabilistic structure with which we shall work, and in rough accordance with decision theory terminology we shall call it a lottery (Figure 3.1). We shall not be concerned about any value assignment to the outcomes, only with their probabilities. The probability of the j^{th} outcome will be denoted p_j, $j = 1, \dots , N$.

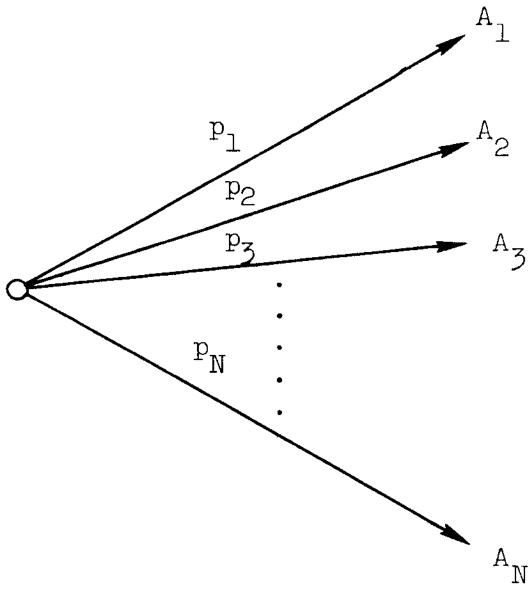


Figure 3.1: A Lottery

We would like to develop a function to measure the "ignorance" or "lack of information" that is expressed in the lottery. We assume that our measure will be totally independent of the values assigned to the outcomes.* It is also to be stressed here that no decisions are being made; we are simply looking for a means of gaining insight into various probabilistic structures.

We wish to develop a function H defined on lotteries that will yield a real number which we may interpret as a measure of the "ignorance" expressed in the lottery. This function will depend only on the

* Implicitly we are assuming Savage's [75] postulate P_4, the substitution principle, that the probabilistic structure is independent of the value of the outcomes. (This assumption is implied by the acceptance of the basic desideratum that the state of information should specify the probability distribution.)

probabilistic structure, e.g., the probabilities $p_1, \dots, p_N$.

Property 0: H is a real-valued function defined on discrete probability distributions $p_1, \dots, p_N$ (i.e., $p_i \geq 0$, $i = 1, \dots, N$, and $\sum_{i=1}^{N} p_i = 1$).

What properties shall we require of the function $H(p_1, \dots, p_N)$? First, it seems reasonable to assert a version of the invariance principle; H should depend on the values of the p_i's, but not on their ordering; rearranging the labels on the outcomes in the lottery leaves H unchanged. We can summarize this assumption as

Property 1: $H(p_1, \dots, p_N)$ is a symmetric function of its arguments.

A second property that seems desirable is that small changes in the probabilities should not cause large changes in H:

Property 2: $H(p_1, \dots, p_N)$ is a continuous function of the probabilities p_i.

The usefulness of the H measure is severely limited unless we can extend it to more complex probabilistic structures than the simple lottery of Figure 3.1. In particular, we should be able to extend H to compound lotteries. It seems natural to require the usual relation for conditional probabilities:

Property 3a: In evaluating the information content of compound lotteries the multiplication law for conditional probabilities holds: e.g., for two events A and B,

$$p(AB) = p(A|B)p(B) .$$

This property is a consistency requirement: The information content should be the same whether the uncertainty is resolved in several stages or all at once. The property (3a) is equivalent to the decomposability or "no fun in gambling" axiom of Von Neumann-Morgenstern utility theory (Luce and Raiffa [52], p. 26).

What form could we assume for H that would allow this measure to be extended to compound lotteries? Perhaps the simplest assumption that we might make is that H takes the form of an expectation:

Property 3b: We may evaluate the information content of a lottery as an expected value over the possible outcomes.

We shall now show that these properties define a specific measure of uncertainty, unique up to a multiplicative constant. The two parts of property 3 essentially determine the form of the uncertainty measure H.

Let us consider the simplest case, $N = 2$. We require that $p_1 + p_2 = 1$, or, equivalently, $p_2 = 1 - p_1$. Suppose the first outcome A_1 occurs, then relative to the original lottery we have gained an amount of information I_1. If A_2 occurs, we gain an amount of information I_2. I_1 and I_2 are now completely arbitrary functions; they represent the amount by which our "ignorance" has been diminished by the resolution of the uncertainty expressed in the lottery. Before it is known whether A_1 or A_2 occurred, then, the expected decrease in "ignorance" is just the quantity we shall define as H in keeping with the expectation property that we have assumed:

$$H(p_1,p_2) = p_1I_1 + p_2I_2 \ .$$

Since p_1 and p_2 are related by $p_2 = 1 - p_1$, and since H is symmetric in its arguments by property 1, I_1 and I_2 must be the same function:

$$\begin{aligned} H(p_1,1-p_1) &= p_1I_1(p_1) + (1-p_1)I_2(1-p_1) \\ &= p_1I(p_1) + (1-p_1)I(1-p_1) \end{aligned} \tag{3.1}$$

because in the simple $N = 2$ case, the probability distribution has only one free parameter, which we may take as p_1.

Now let us consider a more complicated lottery with three distinct outcomes, A_1, A_2, A_3. From the expectation property, we can write

$$\begin{aligned} H(p_1,p_2,p_3) = {} & p_1I_1(p_1,p_2,p_3) + p_2I_2(p_1,p_2,p_3) \\ & + p_3I_3(p_1,p_2,p_3) \ . \end{aligned} \tag{3.2}$$

But <u>if</u> A_1 occurs, the relative probabilities of A_2, A_3 become irrelevant. It should not matter whether these events are considered separately or together as the complement of A_1; I_1 should depend only on p_1 and $1 - p_1$, or $I_1 = I_1(p_1)$. Property 1 implies that I_1, I_2, and I_3 should be the same function, so we can drop the subscripts:

$$H(p_1,p_2,p_3) = p_1I(p_1) + p_2I(p_2) + p_3I(p_3) \ . \tag{3.3}$$

We can then interpret the expectation property (3b) as follows. An "information" random variable is defined on the outcome space by the function I; this is an unusual random variable in that its value depends on the probability measure attached to the individual outcomes. The function H is simply the expected value of this random variable.

The consistency requirement (property 3a) for the evaluation of compound lotteries determines the form of I. Consider the following equivalent lotteries having three possible outcomes:

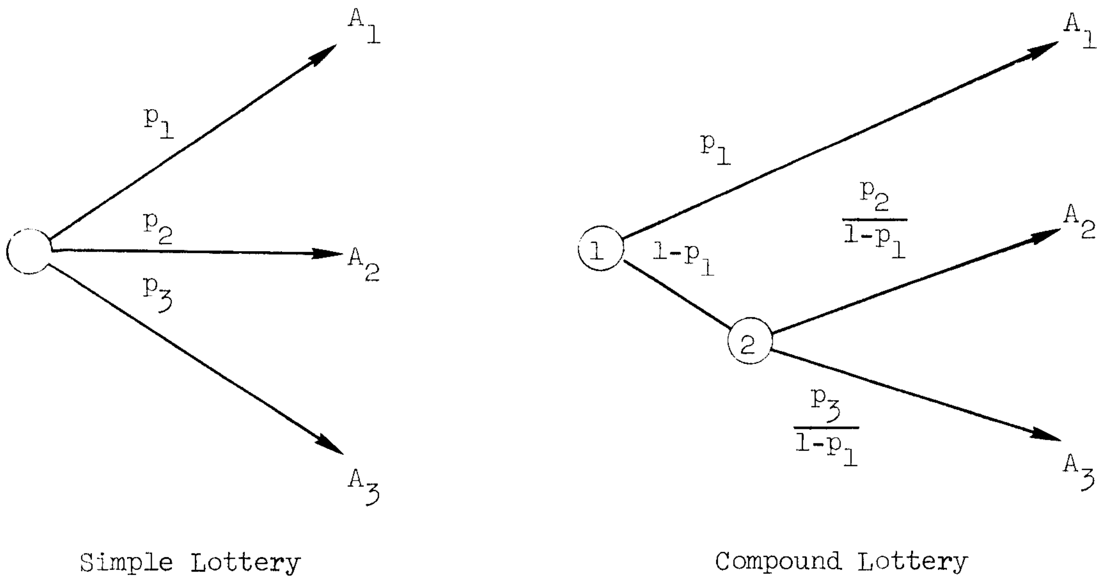


Figure 3.2: Equivalent Lotteries

Suppose that we learn the result of the first chance node in the compound lottery, but not the second. From our calculations on the lottery with two outcomes, the expected gain in information is

$$H(p_1,1-p_1) = p_1 I(p_1) + (1-p_1)I(1-p_1) \ . \tag{3.4}$$

Information from the second chance node will only be relevant if A_1 does not occur, and the probability that A_1 does not occur is $1 - p_1$. From considering the simple lottery we see from equation (3.3) that the expected gain in information from resolving the uncertainty at both chance nodes is

$$\begin{aligned} H(p_1,p_2,p_3) &= p_1 I(p_1) + p_2 I(p_2) + p_3 I(p_3) \\ &= p_1 I(p_1) + (1-p_1)I(1-p_1) \\ &\qquad + (1-p_1)\left[\frac{p_2}{1-p_1} I(p_2) + \frac{p_3}{1-p_1} I(p_3) - I(1-p_1)\right] \\ &= H(p_1,1-p_1) + (1-p_1)\left[\frac{p_2}{1-p_1}\Big(I(p_2) - I(1-p_1)\Big)\right. \\ &\qquad \left. + \frac{p_3}{1-p_1}\Big(I(p_3) - I(1-p_1)\Big)\right] \end{aligned} \tag{3.5}$$

since $1 - p_1 = p_2 + p_3$. We see that the consistency requirement (3a) that the information measure should not depend on whether the uncertainty is resolved all at once or in several stages dictates that the information measure should have an additive form. The first term in the sum represents the expected gain in information from the first chance node, and the second term represents the expected gain in information at the second chance node multiplied by the probability that this node will be reached. The second chance node gives us the outcome A_2 with probability $p_2/(1-p_1)$ and the outcome A_3 with probability $p_3/(1-p_1)$, and using the result (3.1) for a two-outcome lottery, the expected gain in information at this node must be

$$H\left(\frac{p_2}{1 - p_1}, \frac{p_3}{1 - p_1}\right) = \frac{p_2}{1 - p_1} I\left(\frac{p_2}{1 - p_1}\right) + \frac{p_3}{1 - p_1} I\left(\frac{p_3}{1 - p_1}\right) . \qquad (3.6)$$

Comparing this result to (3.5), we find that for these two expressions to be consistent, we require

$$I(p_2) - I(1-p_1) = I\left(\frac{p_2}{1 - p_1}\right)$$

or (3.7)

$$I(p_2) = I(1-p_1) + I\left(\frac{p_2}{1 - p_1}\right) .$$

The solution to this functional equation for arbitrary p_1, p_2 is

$$I(p) = - k \log p$$

where k is conventionally taken to be positive so that $I(p)$ is taken to be an increasing function of $1 - p$.* This convention corresponds to the intuitive notion that the more probable we think it is that an event will not occur, the more information we obtain if it does occur.

Using this result for $I(p)$, we have from (3.3) that the information measure is

$$H(p_1, \ldots , p_N) = - k \sum_{i=1}^{N} p_i \log p_i \qquad (3.8)$$

where k is an arbitrary constant, equivalent to choosing a particular base for the logarithms.** Our derivation has been for the case $N = 3$,

* Details of the solution may be found, for example, in Cox [9], pp. 37-8.

** We shall use natural logarithms unless otherwise specified.

but it is obvious that by induction we could establish the result for arbitrary N. The information measure of a discrete probability distribution specified by expression (3.8) is called the entropy.

The above argument, based in part on a discussion in Feinstein [16], is the reverse of the axiomatic derivation as given originally by Shannon [78] and with some slight modifications by others (Jaynes [33], [35], Khinchin [45], Feinstein [16]). In these derivations the additive property of entropy is taken to be the fundamental assumption rather than the consistency requirement (3a) and the expectation property (3b). Besides properties (0), (1), and (2) an additivity property is required: The information measure of a compound lottery is obtained by adding the information measure at the first node to the sum of the products of the information measures of each successor node and the probability of the branch leading to that node. Suppose we can form a compound lottery with two or more chance nodes by grouping m_i outcomes together and summing the corresponding probabilities q_{ij} to get the probability p_i that one of these outcomes occurred. The additivity assumption is

<u>Property 4</u>: If $p_i = \sum_{j=1}^{m_i} q_{ij} > 0$ for $i \in I$ where I contains at least one integer from $(1, \ldots, N)$, then

$$H(p_1,p_2, \ldots , q_{i1}, \ldots , q_{im_i}, \ldots , p_N) = H(p_1, \ldots , p_N) + \sum_{i \in I} p_i H\left(\frac{q_{i1}}{p_i} , \ldots , \frac{q_{im_i}}{p_i}\right) . \tag{3.9}$$

Feinstein [16] shows that properties (0), (1), (2), and (4) determine

(3.8). Other derivations have employed a fifth assumption which fixes the sign of k and eliminates the need for some of the tortuous mathematics of Feinstein's proof:

> Property 5: $H(\frac{1}{n}, \dots , \frac{1}{n}) = A(n)$ is an increasing function of n.

Shannon's derivation of (3.8) was of this latter form.

Khinchine [45] introduces the additivity assumption, property (4), by saying it is "natural" that the information given by the resolution of uncertainty for two independent lotteries taken together to be the sum of the information given by the resolution of uncertainty for each separately. Why not use some other binary operation such as multiplication to combine the information from independent lotteries? We could start a list of the properties we would like this binary operation to have:

(1) commutative law: It does not matter which lottery is resolved first and which second.

(2) associative law: We should be able to group independent lotteries arbitrarily.

(3) existence of identity element: the lottery with only one (certain) outcome.

and so forth. It is doubtful that a list could be drawn up that would uniquely specify addition as the required binary operation. The crucial determining factor comes in only when we consider compound lotteries. It is not clear that other binary operations than addition could be extended to the dependent case in a meaningful way. Viewing the expectation property (3b) rather than additivity as the fundamental assumption seems to give more insight into why the information measure should have the form (3.8).

We have now developed a measure of the information contained in a probability distribution assigned to an outcome space of N discrete states. The measure is totally divorced from the economic or decision-making aspects of the problem. It is sensitive to the assumed outcome space, and the splitting of any of the outcomes into two or more "sub-states" will cause it to change.

How might this measure help us in assigning probabilistic structures consistent with particular states of information? As an answer, let us note that (3.8) attains a unique maximum for the probability distribution $p_i = 1/N$, $i = 1, \ldots, N$. We have discussed this distribution in the last section, and noted that it corresponded to a state of information of "maximum ignorance," for which even a random relabeling of the states in the outcome space leaves the state of information unchanged. So if we look for the maximum of the entropy function--the probability distribution that will correspond to the largest gain in information when the uncertainty is resolved--we get the same answer as before using the criterion of insufficient reason.

The entropy measure permits us to evaluate and compare the information content of probability distributions. But since the form of the entropy measure depends critically on the expectation property (3b) or an equivalent additivity assumption, the methodology retains an element of arbitrariness that is uncomfortable. Any continuous concave function that is symmetric in its arguments would satisfy properties (0), (1), and (2). We have not established an intuitive justification other than simplicity for assuming the expectation property. As a result we have not shown that the entropy measure provides the only meaningful way to compare probabilistic structures. In fact, in the context of a decision problem we clearly wish to use other measures to evaluate information.*

3.2 The Maximum Entropy Principle

Suppose we consider the following procedure for assigning probability distributions. Since we never wish to let assumptions enter into our analysis without specifically introducing them, let us write down everything we know about the probability distribution. If several distributions are consistent with the information that we have specified, we shall choose that distribution for which the entropy function (3.8) is largest. This criterion is the maximum entropy principle (Jaynes [33]):

> The probability distribution to be assigned on the basis of given information is that distribution whose entropy function is greatest among all those distributions consistent with the information.

In order to use this principle we shall need an operational procedure for specifying information. This is not a simple matter, for

* Given values assigned to each outcome, we shall wish to compare the expected value of taking the best action for each probabilistic structure. This procedure leads to information value theory (Howard [29]). The relation between entropy (3.8) and more general concave functions in information value theory has been explored by DeGroot [11] and Marschak and Miyasawa [54].

often information is only available in a vague form that would not appear suited for translation into quantitative terms. Jaynes [40], after stating the basic desideratum that information should determine probability distributions, restricts his consideration to information that he calls "testable." A piece of information is testable, if given any proposed probability distribution, it is possible to determine unambiguously whether the information agrees with this distribution.

Testable information can be divided into two classes, information concerned solely with the probabilistic structure (the probabilities attached to points in the outcome space), and information concerning values attached to the outcomes. Information in this second class usually takes the form of an equation or inequality involving the expectation of a random variable. In order to have information of the second class one must therefore have a random variable defined (by assigning a numerical value to each point in the outcome space). Information in the first class can be stated as an equation or inequality involving only the probabilities assigned to outcome points; it is not necessary to define a random variable (although one might wish to do so for reasons of convenience).

The difference between these two types of information is subtle, but it provides important insight to the applicability of the maximum entropy principle. If we restrict our consideration to information of the first class (which we shall call probability information), the maximum entropy principle leads to rather trivial results. The second class of information (expectation information) is more interesting, but it is harder to justify how such information might arise in a practical situation.

Let us continue to restrict our attention to uncertain situations with N possible outcomes. Testable information in the first class (probability information) will be composed of equality or inequality statements such as the following:

(a) $p_1 + p_2 + p_3 \leq 0.3$

(b) $p_5 = 0.06$

(c) $\dfrac{k_2 p_2}{\sum_{i=1}^{N} k_i p_i} = 0.7$ where the k_i are a known set of non-negative numbers for $i = 1, \ldots, N$

(d) $\cos^{-1}(p_3) = \pi/4$. (3.10)

Most procedures for encoding probability assignments develop constraints of the types (a) and (b). The subject himself must assimilate his relevant information and process it into testable form. Of course, it may be desirable to formalize this process by constructing a model that relates the probability assignment in question to other uncertain factors, then encoding probability assignments for these.

Constraints of the type (c) might arise in using Bayes' Rule to determine prior probabilities that correspond to a subject's posterior probability assignments. More complicated equations involving the p_i such as (d) are difficult to justify intuitively but still constitute testable information of the probability type.

When testable information has been provided in the form of such probability statements, application of the maximum entropy principle constitutes the following optimization problem:

Choose the probability distribution $(p_1, \ldots, p_N)$ that maximizes

$$- \sum_{i=1}^{N} p_i \log p_i \tag{3.11}$$

subject to the constraints such as (a), (b), (c), (d) that represent the testable information. In addition, of course, the constraints

$$p_i \geq 0, \quad i = 1, \ldots, N \tag{3.12}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{3.13}$$

are needed to insure that $(p_1, \ldots, p_N)$ will be a probability distribution. This formulation is readily extended to include expectation information as well, as we shall see in Chapter 5.

Let us consider how this procedure might apply in a typical situation in which a prior distribution is being encoded. The subject is asked a number of questions regarding his preferences between two lotteries having the same prizes but different probabilistic structures. (For example, see North [60].) The answers determine equations of the form (3.10a,b); it is the encoding process that places the subject's information into testable form.

The encoding process is typically continued until there are enough equations to determine the distribution. In fact, more than this number of equations are usually developed in order to have a check on the subject's consistency. The optimization procedure implied by the principle of maximum entropy is then trivial, because the constraints imply that only one distribution $(p_1, \ldots, p_N)$ is a feasible solution to the optimization problem. For this frequently encountered case the principle of maximum entropy is consistent with the encoding procedure but trivial. Maximizing any other function of the p_i's would lead to the same solution. If the subject has enough information that he can express in testable form to specify a unique probability distribution, there is no need to invoke the principle of maximum entropy. The real gist of the encoding problem lies elsewhere: How does one summarize information into testable form?

Now let us examine the other form of testable information, expectation statements. It is this form of information that has been extensively investigated by the advocates of the maximum entropy principle, but little attention has been devoted to the matter of how such information might arise. Two possibilities suggest themselves: (1) The expectation of a random variable may be known without being derived from probability statements about the individual outcomes. That is, it is possible to know _a priori_ the expectation of a random variable without knowing its distribution function. (2) Knowledge of expectations arises from a series of measurements of similar phenomena: we are told an "average value" without having access to the measurements of the individual instances.

It is this latter type of expectation information that appears in the examples in the literature (Jaynes [35], Chapter 4; the widget problem: Jaynes [39], Tribus and Fitts [88]). Expectation knowledge of the first type, i.e., direct _a priori_ knowledge of an expectation, seems highly unlikely to arise except in situations where a great deal is known about the probability structure. The expectation is a summation over the outcomes of a value attached to each outcome weighted by the probability that the outcome occurs; it is a derived rather than a fundamental concept. Although the probability measure can in some instances be derived from knowledge of expectations (e.g., the characteristic function) it is difficult to ascribe an intuitive meaning to "expectation" except as the average of a large number of identical, independent trials, i.e., a long run average. In any other situation it would seem preferable to encode information using the probability measure directly.

We conclude, therefore, that the maximum entropy principle is not very interesting except in the special case where we are dealing with sequences of identical, independent experiments and our knowledge concerns long-run experimental averages. Further, the foundation for the maximum entropy principle is not as strong as we might like, for neither the assumption of additivity or the expectation property seems clearly intuitive. We shall see in Chapter 5 that the difficulty may be resolved by using the basic desideratum as a starting point rather than the axiomatic derivation presented in this chapter.

Chapter IV

THE MAXIMUM ENTROPY PRINCIPLE IN STATISTICAL MECHANICS: A RE-EXAMINATION OF THE METHODS OF J. WILLARD GIBBS

Entropy considerations did not originate with Shannon's papers. J. Willard Gibbs stated the maximum entropy principle nearly fifty years earlier ([23], pp. 143-144). In many ways Gibbs' development of the entropy principle is more revealing than the contemporary arguments discussed in the last section. The maximum entropy principle may be derived as a direct consequence of the basic desideratum from an invariance to randomization over time. This derivation provides a foundation for statistical mechanics that eliminates any need for an ergodic hypothesis that time averages are equal to expectations over probability distributions. However, the development is so fragmentary that one wonders if Gibbs himself realized the full potential of his methods.* Although the arguments to be presented in this section are drawn from Gibbs' work, their synthesis as a derivation for the maximum entropy principle does not appear to have been previously noted.

4.1 The Problem of Statistical Mechanics

An understanding of Gibbs' reasoning requires some background on the problem in physics with which Gibbs was concerned. Newtonian

* Jaynes [38] has suggested that Gibbs did not live long enough to complete the formulation of his ideas.

mechanics provided the foundation for the physics of the nineteenth century. The philosophical consequence of this viewpoint was a belief in a deterministic universe. Laplace [47] summarized the reasoning through a clever artiface: a superior intellect, the "Laplace demon." The demon could calculate exactly what the future course of the universe would be from Newton's laws, if he were given precise knowledge of the positions and momenta of all particles at any one instant.

It is useful to note the relation of these ideas to Laplace's conception of probability. Such determinism is incompatible with the usual conception of games of chance. The outcome of a throw of the dice is completely determined by the initial conditions; it is a problem in mechanics. If we use probability theory as a means of reasoning about dice, there is no reason we should not use it to reason about any other uncertain occurrences in the physical world. Probability theory was for Laplace a means of making inferences about what is not known but is assumed to be knowable.

The use of probability theory allowed Gibbs to sidestep the need for the demon in applying Newton's laws to the large number of particles in a macroscopic system. This method allowed him to use the laws of mechanics to provide a foundation for the empirical science of thermodynamics. Gibbs' achievement has been acknowledged as one of the great milestones in the history of science, even though the details of his reasoning have been widely misunderstood. The elegance of Gibbs' reasoning becomes apparent when we accept the determinism of the demon analogy and Laplace's interpretation of probability.[*]

[*] The concept of the ensemble is not an intrinsic part of the

We shall now present a derivation of the maximum entropy principle in statistical mechanics. From this maximum entropy principle virtually the entire formalism of statistical mechanics may be easily derived (Jaynes [33], [34], [36], [38]; Tribus [82], [84]). An outline of the argument is as follows: Hamilton's equations are used to describe the dynamic behavior of a system composed of n interacting particles. The uncertainty in the initial conditions for these differential equations of motion is represented by a probability distribution. Liouville's theorem (Gibbs' principle of conservation of probability of phase) provides an important characterization of the evolution of this probability distribution over time; a simple corollary to Liouville's theorem shows that the entropy functional of this probability distribution is constant in time. We then consider the notion of statistical equilibrium which we shall formulate in terms of the basic desideratum: The probability distribution on the initial conditions shall be invariant to a randomization of the time at which these initial conditions are determined. A well-known inequality relation for the entropy functional provides the crucial step in the reasoning: if the probability

argument. Gibbs regarded the ensemble as a means of formalizing the use of probability, and he went to some effort to demonstrate that the fundamental relation of conservation of probability of phase (Liouville's theorem) can be derived without reference to an ensemble: ([23], p. 17) "The application of this principle (conservation of probability of phase) is not limited to cases in which there is a formal and explicit reference to an ensemble of systems. Yet the conception of such an ensemble may serve to give precision to notions of probability. It is in fact customary in the discussion of probability to describe anything which is imperfectly known as something taken at random from a great number of things which are completely described." For the sake of clarity we shall avoid the use of ensembles until the next chapter, where we shall relate them to de Finetti's work on exchangeable sequences.

distribution on the initial conditions is chosen to maximize the entropy functional subject to constraints on the constants of motion, this distribution will remain stationary over time. Gibbs' canonical and microcanonical distributions can be derived from particularly simple constraints on the constants of motion. By this method the entire framework of Gibbs' statistical mechanics is built up from the basic desideratum by means of an invariance principle. We shall now present the derivation in detail.

We shall consider a system composed of n particles governed by the laws of classical mechanics. The particles may interact through forces that depend on the positions of the particles, but not their velocities. External forces may also be considered; we shall not do so here. Hamilton's equations of motion will be used as the formulation of the laws of classical mechanics.* We shall assume that the location of the particles is specified by a set of $3n$ generalized co-ordinates $q_1, \dots, q_{3n}$. The forces affecting the particles are summarized in a Hamiltonian function $\mathcal{H}(p_1, \dots, p_{3n}, q_1, \dots, q_{3n}, t)$ where the p_i are the canonical momenta conjugate to the position co-ordinates q_i. The dynamic behavior of the system is then given by Hamilton's equations:

$$\frac{dp_i}{dt} = \dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} \qquad i = 1, \dots, 3n \tag{4.1}$$

$$\frac{dq_i}{dt} = \dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i} \qquad i = 1, \dots, 3n . \tag{4.2}$$

* The reader who is not familiar with the Hamilton formulation may wish to consult a standard text on mechanics, such as Goldstein [26].

We may represent the state of the system as a point in a $6n$ dimensional phase space Γ whose co-ordinates are $p_1, \ldots, p_{3n} = \vec{p}$, $q_1, \ldots, q_{3n} = \vec{q}$. Given that the system is initially in a state $\vec{p}(t_o)$, $\vec{q}(t_o)$ at time t_o, its motion for all other time is determined by Hamilton's equations, and we may think of the behavior of the system over time as tracting out a trajectory $\vec{p}(t), \vec{q}(t)$ in phase space. Since the solution of Hamilton's equations is unique, these trajectories can never cross each other. A trajectory in phase space must be either a closed curve or a curve that never intersects itself.

Let us consider a volume Ω of phase space at a time t. Consider an arbitrary point $\vec{p}(t), \vec{q}(t)$ in Ω: at another time t' the corresponding location of the system in phase space will be $\vec{p}'(t'), \vec{q}'(t')$. Let us look at the transformation of an infinitesimal volume element in going from the representation at time t to the representation at time t':

$$dp_1 \cdots dp_{3n}\, dq_1 \cdots dq_{3n} = J dp_1' \cdots dp_{3n}'\, dq_1' \cdots dq_{3n}'$$

where J is the Jacobian determinant of the transformation. It is a straightforward matter to show from Hamilton's equations of motion that this determinant is constant in time and equal to one. The proof is given in Gibbs [23], pp. 14-15, or alternatively in many modern texts (e.g., Goldstein [26]). Since trajectories cannot cross, the set of points $\vec{p}(t), \vec{q}(t)$ on boundary of Ω will transform to a set of points $\vec{p}'(t'), \vec{q}'(t')$ that bound a new volume Ω' in phase space. The fact that volume elements are invariant under the transformation from the t representation to the t' representation means that Ω and Ω' must have equal volumes in phase space.

4.2 Probability Distributions on Initial Conditions; Liouville's Theorem

Let us now consider the situation in which the initial conditions are not known. At a particular time t we are uncertain about the location of the system in phase space. We might hypothesize the existence of a "demonic experiment": a device that can measure simultaneously the $3n$ position co-ordinates and the $3n$ momenta needed to locate the system exactly in phase space. We may assign a probability distribution to the outcome of such an experiment performed at a time t_o. $P(\vec{p},\vec{q},t_o)$ will denote the probability that the system is in an infinitesimal volume in phase space containing the point $\vec{p},\vec{q}$ at time t_o. We shall assume that this probability density function exists and is continuous as a function of $\vec{p},\vec{q}$.

From the transformations in phase space determined by Hamilton's equations we can determine the probability distribution over phase space at any arbitrary time t, given the probability distribution at a particular time t_o. We have established that differential volume elements in phase space are invariant under such transformations. Consider the probability that the system will be found in a given volume Ω_o in phase space at time t_o. This is

$$\int_{\Omega_o} P(\vec{p},\vec{q},t_o)dp_1 \cdots dp_{3n}dq_1 \cdots dq_{3n} \; . \tag{4.3}$$

Consider another time t, and let Ω be the volume in phase space bounded by the transformed points on the surface of Ω_o. Since the trajectories cannot cross, if the system is in Ω_o at t_o it will also be in Ω at t, and conversely. Hence:

$$\int_{\Omega_o} P(\vec{p},\vec{q},t_o)dp_1 \cdots dq_{3n} = \int_{\Omega} P(\vec{p},\vec{q},t)dp_1' \cdots dq_{3n}' \ . \qquad (4.4)$$

Let us take Ω very small. Then P is locally constant and may be taken outside the integration. The integrals over the volume are equal, so we find that the probability density function $P(\vec{p},\vec{q},t)$ is constant in time for points in phase space that lie along the trajectory determined by Hamilton's equations. This result was called by Gibbs the principle of conservation of probability of phase, and by modern authors Liouville's Theorem.* From now on we shall use this result and write $P(\vec{p}(t),\vec{q}(t))$ as the probability density function. $P(\vec{p},\vec{q},t)$ will mean the probability distribution over fixed co-ordinates in phase space as a function of time.

Information about the state of the n particle system is often not in the form of knowledge about the generalized co-ordinates and momenta, but rather about quantities that are constants of the motion. The strength of the Hamilton formulation of mechanics is that it lends itself to changes of variables, and we can transform to a new set of co-ordinates for phase space that include these constants of the motion.

For a dynamical system of n interacting particles obeying Hamilton's equations there will be in general $6n$ constants of the

* Liouville's Theorem is often stated in terms of a density of points in an ensemble in phase space rather than a probability density, but as Gibbs pointed out, the two formulations are conceptually equivalent.

motion. $6n - 1$ of these specify a trajectory in phase space while the remaining one locates the system on the trajectory at a particular time. Let us denote these constants as $c_1, \cdots, c_{6n}$.

The position of the system in phase space $\vec{p}(t), \vec{q}(t)$ can be specified as a function of these constants $c_1, \cdots, c_{6n}$ and the time t:

$$\begin{aligned} \vec{p}(t) &= \vec{p}(c_1, \cdots, c_{6n}, t) \\ \vec{q}(t) &= \vec{q}(c_1, \cdots, c_{6n}, t) \,. \end{aligned} \tag{4.5}$$

These expressions represent the solution of the $6n$ Hamilton differential equations of motion in terms of the $6n$ constants of integration for these equations and the independent variable t. We can now consider a transformation from the original phase space Γ to a new $6n$ dimensional space whose co-ordinates are $c_1, \cdots, c_{6n}$. A differential volume element in the new space may be related to a volume element in the old space in the usual way using the Jacobian determinant of the transformation equations (4.5):

$$dp_1 \cdots dp_{3n} dq_1 \cdots dq_{3n} = \frac{\partial(p_1,\ldots,p_{3n},q_1,\ldots,q_{3n})}{\partial(c_1,\ldots,c_{6n})} dc_1 \cdots dc_{6n} . \tag{4.6}$$

Since volume elements in phase space remain constant as the system evolves through time, and since a system having constants of motion within specified limits will remain within these limits, we conclude that the Jacobian determinant

$$\frac{\partial(p_1, \cdots, p_{3n}, q_1, \cdots, q_{3n})}{\partial(c_1, \cdots, c_{6n})}$$

must be constant in time.

The implication of this change of variables on Liouville's theorem was clearly evident to Gibbs, and we shall state the following general version of Liouville's theorem in Gibbs' own words ([23], p. 18):

> When the differential equations of motion are exactly known, but the constants of the integral equations imperfectly determined, the coefficient of probability of any phase at any time is equal to the coefficient of probability of the corresponding phase at any other time. By corresponding phases are meant those which are calculated for different times from the same values of the arbitrary constants of the integral equations.

By coefficient of probability of phase Gibbs meant the probability density function $P(\vec{p},\vec{q})$ in phase space. For "phase" read phase space, and for "constants of the integral equations" read constants of integration.

Only a few of the constants of motion will generally be of interest: those that correspond to the conservation laws of physics and are therefore macroscopically observable. The other constants will be highly variable with the precise location of the system in phase space at a reference instant t_o. They are not susceptible to reproducible measurement, and typically there is no way to learn their values without the assistance of Laplace's demon to solve Hamilton's equations. The constant of the motion that is of greatest interest in statistical mechanics is the energy of the system. Other constants that are occasionally considered are the components of total angular momentum (Gibbs [23], p. 38).

Let us assume that there exist I known functions $\theta_1, \theta_2, \ldots, \theta_I$ of the generalized co-ordinates and momenta $\vec{p}(t), \vec{q}(t)$ whose values remain constant as the system evolves over time. The functions θ_i

transform the specification of initial conditions from the phase space $\vec{p}(t)$, $\vec{q}(t)$ to a new set of variables in which the first I variables, $c_1, \dots, c_I$, do not change over time as a consequence of Hamilton's equations.

$$\theta_i\left(\vec{p}(t),\vec{q}(t)\right) = c_i \ , \qquad i = 1, 2, \dots, I \ .$$

The quantities c_i are presumed to be measurable, but lack of information will cause these values to be uncertain. We shall suppose that knowledge about the constants c_i may be encoded as a probability distribution over the possible values of the c_i that would result from a measurement of infinite precision. This distribution can be represented by the joint cumulative distribution function $F(c_1,c_2, \dots, c_I)$. We now wish to show the following convexity property:

Lemma 1: Suppose $P(\vec{p},\vec{q},t_o)$ and $P'(\vec{p},\vec{q},t_o)$ are two distributions over phase space at t_o that are consistent with a given probability distribution $F(c_1, \dots, c_I)$ on the constants of motion $c_1, \dots, c_I$. That is, the integral of $P(\vec{p},\vec{q},t_o)$ over those regions of phase space such that $\theta_i\left(\vec{p}(t_o),\vec{q}(t_o)\right) \le c_i$, for $i = 1, \dots, I$ is equal to $F(c_1, \dots, c_I)$ for all possible values of $c_1, \dots, c_I$:

$$\int_{\Gamma \text{ such that } \theta_i(\vec{p},\vec{q}) \le c_i, \ i = 1,\dots,I} P(\vec{p}(t_o),\vec{q}(t_o),t_o)dp_1 \cdots dp_{3n}dq_1 \cdots dq_{3n} = F(c_1, \dots, c_I) \quad (4.7)$$

The same relation (4.7) is also assumed to hold for the distribution $P'(p,q,t_o)$. Then the distribution formed by taking a positive linear combination

$$P^*(\vec{p},\vec{q},t_o) = \alpha P(\vec{p},\vec{q},t_o) + (1-\alpha)P'(\vec{p},\vec{q},t_o) \qquad 0 \leq \alpha \leq 1 \tag{4.8}$$

also satisfies the probability constraint (4.7) for all possible values $c_1, \ldots, c_I$ of the constants of motion. Further, the distributions P, P', and P^* satisfy the probability constraint (4.7) for all times t.

Proof: The fact that P^* satisfies (4.7) is a trivial consequence of the fact that integration over phase space is a linear operation. The distributions P, P', P^* satisfy (4.7) for all times t, given that they satisfy (4.7) at t_o, since (4.7) is equivalent to

$$\int_{-\infty}^{c_1} \cdots \int_{-\infty}^{c_I} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(\vec{p}(t),\vec{q}(t))$$

$$\frac{\partial(p_1(t),\ldots,p_{3n}(t),q_1(t),\ldots,q_{3n}(t))}{\partial(c_1,\ldots,c_{6n})} dc_1 \cdots dc_I dc_{I+1} \cdots dc_{6n}$$

$$= F(c_1, \ldots, c_I) \, . \tag{4.9}$$

$P(\vec{p},\vec{q})$ and the Jacobian determinant are constant in time as $\vec{p}(t)$, $\vec{q}(t)$ move in accordance with Hamilton's equations, so if (4.9), hence (4.7) is satisfied at one time t_o, it is satisfied for all times t.

Lemma 2: Suppose that we have a set of probability distributions $P_x(\vec{p},\vec{q})$ over phase space indexed by a parameter x, each of which satisfies (4.7) at a specific time t_o. Suppose $\varphi(x)$ is a probability distribution over the values of the parameter x. Then the composite distribution

$$P^*(\vec{p}(t),\vec{q}(t)) = \int_X P_x(\vec{p}(t),\vec{q}(t))d\varphi(x) \qquad (4.10)$$

satisfies (4.7) for all times t.

Proof: The proof follows immediately from the fact that the integrations over x and over phase space can be interchanged,* and then the application of Liouville's theorem as in lemma 1.

In general the probability density $P(\vec{p}(t),\vec{q}(t))$ that the system will be found in a fixed region of phase space is not constant, but changes in time. The exceptional situation in which $P(\vec{p}(t),\vec{q}(t))$ is constant in time for any fixed point in phase space was defined by Gibbs to be statistical equilibrium. For any arbitrary volume Ω fixed in phase space, the probability that the system will enter the volume in a time Δt is equal to the probability that the system will leave Ω in Δt. The meaning of Liouville's theorem is that $P(\vec{p}(t),\vec{q}(t))$ is constant in time for a region moving through phase space according to Hamilton's equations even if one does not have statistical equilibrium.

Can we summarize Liouville's theorem without relying on a function defined over $6n$ dimensional phase space? Supposing we take the integral of $P(\vec{p},\vec{q})$ over phase space; this integral must be unity if $P(\vec{p},\vec{q})$ is to be a probability density so we learn nothing. But suppose we consider the natural logarithm:

* Fubini's theorem.

$$\eta(\vec{p}(t),\vec{q}(t)) = -\log P(\vec{p}(t),\vec{q}(t)) \ . \qquad (4.11)$$

This function will be constant for points moving in phase space according to Hamilton's equations. Furthermore, consider the expectation of $\eta(\vec{p},\vec{q},t)$ over phase space, which is the entropy functional

$$H(P,t) = -\int_\Gamma P(\vec{p}(t),\vec{q}(t)) \log P(\vec{p}(t),\vec{q}(t)) dp_1 \cdots dq_{3n} \ . \qquad (4.12)$$

Both P and η are constant and the volume element remains constant for each point in phase space when we transform according to Hamilton's equations, and we come to the important consequence of Liouville's theorem:

Gibbs H-Theorem: The function

$$H(P,t) = -\int_\Gamma P(\vec{p}(t),\vec{q}(t)) \log P(\vec{p}(t),\vec{q}(t)) dp_1 \cdots dq_{3n} \qquad (4.13)$$

is constant over time. This result is stated by Gibbs [23], p. 144.

Gibbs associated H with the entropy of a thermodynamic system, and later authors have criticized him severely on this point: "Entropy must increase in an irreversible thermodynamic process, so H cannot be the entropy." For a discussion of the relation between the Gibbs H function, the thermodynamic entropy, and the Second Law of Thermodynamics, the reader is referred to Jaynes [38].* Henceforth we shall refer to the H function as the entropy functional of the probability distribution over phase space.

* The Szilard articles that appeared in Zeitschrift für Physik in the 1920's also yield considerable insight into the relation between the two kinds of entropy. Summaries in English are found in [14] and [89].

4.3 Statistical Equilibrium and the Maximum Entropy Principle

We now use the invariance of H over time to arrive at a powerful characterization of statistical equilibrium. We shall need the following lemma, which appears in Gibbs [23] as Theorem VIII, p. 135.

Lemma 3: Let $P_1, P_2, \ldots$ be a set of probability distributions over phase space. Suppose we consider a random mechanism where the i^{th} distribution P_i is chosen with probability $\varphi(i)$. Denote the composite distribution over phase space

$$\sum_i P_i \varphi(i) = P_o \qquad ; \qquad (4.14)$$

then the entropy functions of these distributions obey the following inequality relation

$$H(P_o) \geq \sum_i H(P_i)\varphi(i) \qquad (4.15)$$

with equality holding if and only if $P_o = P_i$ identically for all $i = 1, 2, \ldots$.

Proof: This lemma is often stated as "the conditional entropy is never greater than the unconditional entropy." (Shannon [78], p. 22, Khinchin [45], p. 36, Feinstein [16], p. 15.) The proof depends on the fact that $f(x) = x \log x$ is a strictly convex function. The version used here is that given by Gibbs [23], pp. 136-7. (Another version is found in Chapter 5 preceding equation (5.46).)

Consider the function

$$Q_i = Q_i(\vec{p},\vec{q}) = P_i \log P_i - P_i \log P_o - P_1 + P_o \, . \qquad (4.16)$$

We shall see that it is always positive except where $P_i = P_o$. Consider any point $\vec{p}, \vec{q}$ in phase space; then P_i and P_o may be regarded as any positive numbers. If P_o is held constant and P_i varied, we have

$$\frac{dQ_i}{dP_i} = \log P_i - \log P_o$$

$$\frac{d^2Q_i}{dP_i} = (P_i)^{-1} .$$

Q_i and dQ_i/dP_i vanish for $P_i = P_o$, and the second derivative is everywhere positive. Hence Q_i must be positive if $P_i \neq P_o$. Further,

$$\sum_i \varphi(i)Q_i > 0$$

unless $P_i = P_o$ for all i. Since

$$\sum_i \varphi(i)Q_i = \sum_i \varphi(i)[P_i \log P_i - P_i \log P_o - P_i + P_o]$$

$$= \sum_i \varphi(i)P_i \log P_i - P_o \log P_o > 0$$

we have shown that for any point $\vec{p}, \vec{q}$ in phase space,

$$- P_o(\vec{p},\vec{q}) \log P_o(\vec{p},\vec{q}) > - \sum_i \varphi(i)P_i(\vec{p},\vec{q}) \log P_i(\vec{p},\vec{q})$$

with equality only if $P_o = P_i$ for all i. Integrating over phase space proves the lemma. (Strictly speaking the equality holds even if $P_o(\vec{p},\vec{q}) \neq P_i(\vec{p},\vec{q})$ providing the latter occurs only on sets of measure zero. We ignore such technicalities.)

There is an immediate corollary to this result.

Lemma 4: The distribution that maximizes the entropy functional H subject to any set of probability constraints of the form (4.7) is unique.

Proof: Suppose the contrary: there exist two distinct distributions P_1 and P_2 that maximize the entropy functional on some outcome space, subject to the constraints.

Consider the composite distribution $\alpha P_1 + (1-\alpha)P_2$, where $0 < \alpha < 1$. By lemma 1 if P_1, P_2 satisfy the constraints, $\alpha P_1 + (1-\alpha)P_2$ satisfies the constraints. But by lemma 3 we have

$$H(\alpha P_1 + (1-\alpha)P_2) > \alpha H(P_1) + (1-\alpha)H(P_2) \ . \qquad (4.17)$$

Hence P_1 and P_2 cannot maximize the entropy functional H subject to the given probability constraints unless $P_1 = P_2$ identically.

We are now ready to derive the main result of this section. The argument is suggested by Gibbs in the last paragraph, [23], p. 151, but it is not clear if Gibbs intended the passage to be the justification for the maximum entropy principle or simply a deduction of the consequences of lemma 3.

Consider a system in equilibrium. We shall characterize equilibrium in the following way: our state of information regarding the outcome of a demonic experiment that would locate the system exactly in phase space is the same, whether we

(1) perform the experiment at a fixed time t_o

(2) perform the experiment at a time selected randomly from a set $t_1, t_2, \ldots$.

We now apply the basic desideratum: if the state of information is the same for both situations (1) and (2), we should assign the same distribution in phase space to the outcome of the experiment for both situations.

Suppose we assign the distribution that maximizes the entropy functional, while being consistent with the probabilistic information that we have about the constants of the motion. Supposing situation (1), this distribution $P(\vec{p},\vec{q},t_o)$ will apply for the demonic experiment performed at a fixed time t_o. Now let us compare this distribution with situation (2). Suppose the experiment will be performed at a time t_i with probability $\varphi(i)$. We use Hamilton's equations to derive $P(\vec{p},\vec{q},t_i)$ for $i = 1, 2, \ldots$. Because of the Gibbs' H theorem each of these distributions will have the same entropy as $P(\vec{p},\vec{q},t_o)$. The $P(\vec{p},\vec{q},t_i)$ also satisfy the same probabilistic constraints on the constants of the motion from the general version of Liouville's theorem. Hence each must be a maximum entropy distribution. By lemma 4 they must be the same distribution, so we have shown that the maximum entropy distribution is a constant over time.

Supposing we assign a distribution $P(\vec{p},\vec{q},t_o)$ that does not maximize the functional H subject to satisfying a set of probabilistic relations on the constants of motion. The Gibbs' H theorem still holds, so

$$H(P(\vec{p},\vec{q},t_i)) = H(P(\vec{p},\vec{q},t_o)) \qquad (4.18)$$

for i = 1, 2, ... But consider the composite distribution:

$$\sum_i \varphi(i)P(\vec{p},\vec{q},t_i) = P*(\vec{p},\vec{q}) \ . \tag{4.19}$$

By the lemma we have

$$H(P*(\vec{p},\vec{q})) \geq \sum_i \varphi(i)H(P(\vec{p},\vec{q},t_i)) \tag{4.20}$$

with equality holding only if

$$P*(\vec{p},\vec{q}) = P(\vec{p},\vec{q},t_i) \tag{4.21}$$

for all i. Now we have no reason to require that equality holds; the composite distribution may give a larger value to the H functional while still satisfying the constraints on the constants of motion. In this case we may have different distributions $P(\vec{p},\vec{q},t_i)$ for each i. But now we invoke the basic desideratum: $P*(\vec{p},\vec{q})$ has a higher value of H than $P(\vec{p},\vec{q},t_o)$, yet these two probability distributions must be identical since they correspond to identical states of information. Only in the case in which $P(\vec{p},\vec{q},t_o) = P(\vec{p},\vec{q},t_i)$ for all i (and recall that the t_i can be chosen arbitrarily) do we have $P*(\vec{p},\vec{q}) = P(\vec{p},\vec{q},t_o)$ as required by the basic desideratum.

We have now achieved a fundamental insight into statistical equilibrium. By invoking an invariance to randomization over time, we find that the probability distribution over phase space has to be constant in time, which is Gibbs' definition of statistical equilibrium. But more important, we have learned how to generate an immense class of distributions that have the property that they remain constant over time: we choose them to have maximum entropy subject to probabilistic constraints on the constants of the motion.

Particularly simple forms for these constraints lead to the well-known distributions in statistical mechanics. If the only probabilistic constraint is on the total energy of the system, and it is presumed that this energy is precisely known, the resulting distribution is called microcanonical. The canonical distribution results from maximization of the entropy functional subject to a constraint on the expected value of the energy over phase space. We shall see how to handle this expected value constraint in the next chapter.

The formulation of statistical mechanics that has just been presented avoids the need for the usual[*] ergodic hypothesis equating expectations over phase space with the time averages that are actually measured in the laboratory. By invoking the basic desideratum for a system in equilibrium we were able to develop a probability distribution over phase space that reflected our knowledge about the state of the system. Hamilton's equations are deterministic, and once the initial conditions are specified, the behavior of the system can be (in principle) predicted for all time. Liouville's theorem and its corollary, the Gibbs'H theorem, are direct consequences of the fact that the Hamilton's equations governing the dynamic evolution of the system introduce no element of uncertainty. It is in fact obvious that the stochastic process constructed by observing the system over time is not metrically transitive and therefore the usual proof of the ergodic theorem does not hold (Loève [51], pp. 423-4).

[*] For example, see [31], p. 203, and [77], p. 9.

4.4 Relation to the Criterion of Insufficient Reason

The preceding material indicates the power and subtlety of the entropy concept. Yet the probability distributions having maximum entropy subject to probability assignments of the type (4.7) to the constants of motion have a simple characterization according to the criterion of insufficient reason. This insight results from a theorem due to Gibbs (Chapter XI, theorem IV [23], p. 132).

Theorem: Suppose probability assignments of the form (4.7) are given on the constants of motion $c_1, \ldots , c_I$. The probability distribution over phase space having maximum entropy while satisfying this information constraint is the one for which $P(\vec{p},\vec{q})$ is a function only of the constants of motion $c_1, \ldots , c_I$.

Proof: We shall use Gibbs' method of proof, which is essentially a calculus of variations argument.

Let $\eta = \eta(c_1, \ldots , c_I) = - \log P(\vec{p},\vec{q}) = - \log P(c_1, \ldots , c_I)$ be a function that depends on the location $\vec{p}, \vec{q}$ in phase space only through the constants of motion $\theta_i(\vec{p},\vec{q}) = c_i$, $i = 1, \ldots , I$. Let $\delta\eta$ be an arbitrary function of phase space, subject only to the conditions that the probability distribution $P*$ over phase space corresponding to $\eta + \delta\eta$:

$$P*(\vec{p},\vec{q}) = e^{-\eta(c_1,\ldots,c_I)-\delta\eta(\vec{p},\vec{q})} \qquad (4.22)$$

satisfies the following conditions. First, it is a probability distribution, and second, it satisfies the information constraint (4.7).

Writing these conditions out,

$$\int_\Gamma e^{-\eta(c_1,\ldots,c_I)-\delta\eta(\vec{p},\vec{q})} dp_1 \cdots dq_{3n} = \int_\Gamma e^{-\eta(c_1,\ldots,c_I)} = 1 \qquad (4.23)$$

$$\int_{\substack{\Gamma \text{ such that} \\ c_i \le \theta_i(\vec{p},\vec{q}) \le c_i + \Delta c_i \\ i = 1, \ldots, I}} e^{-\eta(c_1,\ldots,c_I)-\delta\eta(\vec{p},\vec{q})} dp_1 \cdots dq_{3n} = \int_{\substack{\Gamma \text{ such that} \\ c_i \le \theta_i(\vec{p},\vec{q}) \le c_i + \Delta c_i \\ i = 1, \ldots, I}} e^{-\eta(c_1,\ldots,c_I)} dp_1 \cdots dq_{3n}$$

$$= \frac{d}{dc_1} \cdots \frac{d}{dc_I} F(c_1, \ldots , c_I) \ . \qquad (4.24)$$

Since $\eta(c_1, \ldots , c_I)$ is approximately constant, we can multiply both sides of (4.24) by it to obtain

$$\int_{\substack{\Gamma \text{ such that} \\ c_i \le \theta_i(\vec{p},\vec{q}) \le c_i + \Delta c_i \\ i = 1, \ldots, I}} \eta e^{-\eta-\delta\eta} dp_1 \cdots dq_{3n} = \int_{\substack{\Gamma \text{ such that} \\ c_i \le \theta_i(\vec{p},\vec{q}) \le c_i + \Delta c_i \\ i = 1, \ldots, I}} \eta e^{-\eta} dp_1 \cdots dq_{3n} \ . \qquad (4.25)$$

Now we wish to prove that $P = e^{-\eta}$ is the maximum entropy distribution. This is equivalent to

$$\int_\Gamma (\eta+\delta\eta) e^{-\eta-\delta\eta} dp_1 \cdots dq_{3n} \le \int_\Gamma \eta e^{-\eta} dp_1 \cdots dq_{3n} \qquad (4.26)$$

with equality holding only for $\delta\eta = 0$ for all points $\vec{p}, \vec{q}$ in phase space. Integrating (4.25) over all values of $c_1, \ldots , c_I$, we see that (4.26) is equivalent to

$$\int_\Gamma \delta\eta e^{-\eta-\delta\eta} dp_1 \cdots dq_{3n} \le 0 \qquad (4.27)$$

Using (4.23),.

$$\int_\Gamma \delta\eta e^{-\eta-\delta\eta} dp_1 \cdots dq_{3n}$$
$$= \int_\Gamma e^{-\eta}(\delta\eta e^{-\delta\eta} - 1 + e^{-\delta\eta}) dp_1 \cdots dq_{3n} \ . \qquad (4.28)$$

Consider the function $f(x) = xe^x - e^x + 1$. Since $f'(x) = xe^x$, $f(x)$ attains a unique minimum of zero at $x = 0$. Hence at any point $\vec{p}, \vec{q}$ in phase space,

$$\delta\eta e^{-\delta\eta} - 1 + e^{-\delta\eta} = -f(-\delta\eta) \qquad (4.29)$$

must be negative unless $\delta\eta = 0$. Therefore the integrand in (4.28) is everywhere non-positive, and the integral can only be zero if $\delta\eta(\vec{p},\vec{q}) = 0$ for all points $\vec{p}, \vec{q}$ in phase space.

The insight that the theorem gives is the following. We have a probability distribution $F(c_1, \ldots, c_I)$ on the constants of motion and a series of equations

$$\theta_i(\vec{p},\vec{q}) = c_i \qquad i = 1, \ldots, I$$

relating these constants to positions in phase space. We assign a distribution over phase space such that a change of variables from $\vec{p}, \vec{q}$ to $c_1, \ldots, c_I$ gives us the distribution $F(c_1, \ldots, c_I)$, and such that all points $\vec{p}, \vec{q}$ that generate the same set of values for the constants of the motion are equally probable. We could have obtained this result directly by using the criterion of insufficient reason. We specify information on the constants $c_1, \ldots, c_I$, and assign a uniform distribution over phase space for the remaining $6n - I$ constants of the motion. The power of the method in statistical mechanics depends on the fact that typically there is only one constant of the motion (energy) and the dimensionality of phase space, $6n$, is a huge number, of the order of 10^{24}.

Chapter V

STATISTICAL ENSEMBLES AND THE MAXIMUM ENTROPY PRINCIPLE

In Chapter 3 we examined the maximum entropy principle for assigning probability distributions and found that it lacked an intuitive conceptual foundation. Further, it did not lead to significant insight when information was provided in the form of probability statements; the proponents of this principle have introduced information in the form of statements about the expectation of a random variable. We argued that only when these expectations were equivalent to long run averages did such statements have an intuitive meaning.

In Chapter 4 we examined how the maximum entropy principle arises in statistical mechanics as a consequence of statistical equilibrium. We noted that a probability distribution over phase space that maximizes entropy subject to probabilistic constraints on a subset of the constants of motion is equivalent to a uniform distribution over the constants of motion that do not enter into these constraints. If the energy of the system is known exactly, the resulting distribution is the microcanonical distribution. The canonical or Boltzmann distribution corresponds to knowledge of the ensemble average of the energy. We do not yet know how to deal with this kind of information.

The next two chapters will be devoted to resolving these issues that have been raised in earlier chapters: (1) a better justification for the maximum entropy principle, and (2) a means of treating expectation constraints. We shall turn our attention to situations in which repeated indistinguishable experiments are considered. The probabilities of classical statistics exist only in repetitive situations where they are defined to be equal to long run frequencies. Our Bayesian viewpoint on probability coincides with this classical view in the limit where the long run frequencies are known. The long run frequency distribution is generally not known; rather we are uncertain about it. It is this uncertainty that motivates our work in the next two chapters. Inferring which long run frequency distribution is appropriate for a given repetitive situation is the basic problem with which we shall be concerned.

We shall begin by formalizing the notion of repeated, indistinguishable experimental trials; following Gibbs we shall call a collection of repeated, indistinguishable experiments an ensemble. We may relate an ensemble to the basic desideratum through de Finetti's exchangeability concept. Our state of information about a sequence of experiments is unchanged by any arbitrary permutation in the order of the experiments. An ensemble is then a collection of exchangeable trials, and we shall use the terms interchangeably. A theorem due to de Finetti formalizes the relation between exchangeable sequences and long run frequency distributions.

The remainder of the chapter builds up the mathematical proofs necessary to understand the maximum entropy principle and the significance of information in the form of ensemble averages. In Chapter 6 we shall summarize the implications of these results and examine some possible extensions.

5.1 Statistical Ensembles as Exchangeable Sequences

The concept of an ensemble goes back to the earliest writings on probability. These writings concerned games of chance, for which the notion of identical, independent trials is intuitively obvious. The objective of probability theory was to make deductions about the outcome of these games of chance from the probability law of an underlying random mechanism, such as dice or cards.

The notion of identical, non-interacting experiments might well have been what Gibbs had in mind in postulating an ensemble of systems as the basis for statistical mechanics: Gibbs defines an ensemble on the opening page of his preface as follows ([23], p. vii):

> We may imagine a great number of systems of the same nature, but differing in the configurations and velocities which they have at a given instant, and differing not merely infinitesimally, but it may be so as to embrace every conceivable combination of configuration and velocities.

It seems possible that Gibbs was thinking of a great many "demonic experiments" to measure the configurations and velocities of the particles composing the system. Systems "of the same nature" would imply that these experiments are run on systems that are indistinguishable to the physicist in terms of the (macroscopic) measurements that he is able to make.

The interpretation expressed above is not the one currently prevailing in physics. Gibbs' definition is taken quite literally: An ensemble is viewed as a large number of "mental copies" of the real physical system under consideration ([31],[77]). These "copies" are to occupy every conceivable position in phase space consistent with the macroscopic "state" of the system. We showed in the last section that many of Gibbs' results could be developed by using a probability distribution on the initial conditions (i.e., the position in phase space at time t_o) instead of the "density function of the ensemble in phase space" that is usually employed in deriving these results.

This prevailing interpretation of ensemble leads to many conceptual problems. The Gibbs ensemble has no physical reality, but physicists have treated it as if it did. The basis for choosing an ensemble distribution, the meaning of an "ensemble average," and the need for an ergodic hypothesis are points that have bothered most writers on statistical mechanics. Typically these difficulties are passed over with an apology and a pragmatic justification: Statistical mechanics is useful because it gives the right answers. ([31],[77].)

The conception of an ensemble as a collection of repeated indistinguishable experiments allows these difficulties to be resolved or avoided. We shall take ensemble to mean a collection of experiments without attaching a probability law to the experimental outcomes. We might stress that this interpretation is <u>not</u> standard, either in physics or in probability and statistics.*

* The difficulty inherent in conceiving of the ensemble as a physical entity becomes manifest when one tries to ascribe a meaning to the ensemble average energy. The usual procedure in physics follows a suggestion made by Einstein in 1914. (For a summary of the argument in English, see von Neumann [89], p. 361 ff.) Consider n non-interacting replicas of the system all placed in a heat bath and allowed to come to thermal equilibrium (by exchanging energy). The distribution of energy among the n replicas may be then shown to follow the Boltzmann distribution by exactly the same methods Boltzmann applied to the molecules of an ideal gas. In the formulation we suggest, the ensemble is composed of n replicas of an experiment, conducted at the same temperature. The concept of temperature implies that heat (energy) would not spontaneously flow from one system to another system at the same temperature

A basis for assigning probabilities to a collection of repeated, indistinguishable experiments was given more than thirty years ago by de Finetti. He introduced the notion of exchangeability, as follows ([20], p. 123):

> We shall say that $x_1, x_2, \ldots, x_n, \ldots$ are exchangeable random quantities if they play a symmetrical role in relation to all problems of probability, or, in other words, if the probability that $x_{k_1}, x_{k_2}, \ldots, x_{k_n}$ satisfy a given condition is always the same however the distinct indices $k_1 \cdots k_n$ are chosen.

We can regard exchangeability as an invariance concept: Exchangeability can be interpreted as an application of the basic desideratum. If we have a sequence of uncertain quantities (experimental outcomes), $x_1, \ldots, x_n$, and our state of information is not changed by permuting or even randomly assigning the labels specifying the order of the quantities in the sequence, then the quantities are exchangeable. An ensemble composed of a collection of indistinguishable experiments is then a collection of exchangeable random quantities.

Exchangeability is a concept that applies to a state of information. By defining an ensemble as a sequence of exchangeable

if the two systems were brought into thermal contact at some particular time and place; however, the experiments might take place at different times using the same system or different systems. Our formulation is consistent with Einstein's conception, but free of the conceptual difficulties imposed by a heat bath filled with mental copies. (For an example of the confusion these conceptual difficulties can engender, the reader is referred to Schrödinger [77], p. 3.)

The concept of an ensemble in statistics is usually used in reference to a stochastic process to mean the set of possible outcomes of the process over time together with the probability law assigned to these outcomes ([10], p. 39; [62], p. 72).

experiments, we emphasize that an ensemble is a mental concept that may have no actual physical counterpart. While we might think about a large number of indistinguishable experiments, it may be possible to perform only one experiment.

Nothing has been said about the timing of the experiments. They might be performed simultaneously on experimental systems that are judged indistinguishable, or they might be repeated experiments using the same system. In this latter case of sequential experiments over time, exchangeability is equivalent to stationarity and independence assumptions: the same probability distribution is assigned to the experiment no matter when it is performed or what sequence of outcomes preceded it.

5.2 De Finetti's Theorem

Viewing ensembles as exchangeable sequences allows a very important theorem due to de Finetti to be applied. A rather lengthy proof is found in de Finetti's original paper [20]; a more elegant but less accessible version is available in Loève [51].

De Finetti's Theorem: Let $x_1, x_2, \ldots, x_n$ be a sequence of exchangeable random quantities. Then $x_1, \ldots, x_n$ are conditionally independent with a common probability distribution. That is, the joint distribution of any subset $x_{k_1}, x_{k_2}, \ldots, x_{k_j}$ of the random quantities may be written as an expectation of the product of conditional distributions*

* From this point on we shall make extensive use of inferential notation. (see, for example, Howard [30].) The brackets $\{\ \mid \mathcal{S}\}$ denote a probability mass function for discrete outcome spaces, or a probability density function for continuous spaces, assigned conditionally on a particular state of information $\mathcal{S}$ · $\int_x$ represents a generalized

$$\{x_{k_1}, \ldots , x_{k_j} | \mathcal{E}\} = \int_{\omega} [\{x_{k_1} | \omega, \mathcal{E}\}\{x_{k_2} | \omega, \mathcal{E}\} \cdots \{x_{k_j} | \omega, \mathcal{E}\}]\{\omega | \mathcal{E}\} . \tag{5.1}$$

The distributions $\{x_{k_j} | \omega, \mathcal{E}\}$ are identical, and ω is a parameter indexing possible distributions.

De Finetti's theorem has a straightforward interpretation in the framework of Bayesian inference. Suppose there is a process generating exchangeable random quantities $x_1, x_2, \ldots , x_n$. Initially we have a prior probability assignment $\{\omega | \mathcal{E}\}$ on the frequency distribution or limiting histogram ω that would summarize the observed outcomes in a large number of exchangeable experimental trials. The prior probability assigned to the outcomes of the first j trials is, from (5.1),

$$\{x_1, \ldots, x_n | \mathcal{E}\} = \int_{\omega} [\{x_1 | \omega, \mathcal{E}\}\{x_2 | \omega, \mathcal{E}\} \cdots \{x_n | \omega, \mathcal{E}\}]\{\omega | \mathcal{E}\} \tag{5.2}$$

and after learning the results $x_1, \ldots , x_n$, we could use these to update the probability assigned to x_{n+1}:

summation over the outcome space of the random quantity x. The expectation of a function of a random quantity x will be denoted

$$< \varphi(x) | \mathcal{S} >$$

where $\mathcal{S}$ is the state of information on which the probability distribution of x has been assigned. In some cases we shall wish to switch to a functional notation by defining, for example,

$$p(x) = \{x | \omega, \mathcal{E}\} .$$

$\mathcal{E}$ is conventionally used to represent the "prior" information that is available at the beginning of the analysis.

$$\{x_{n+1}|x_1,x_2,\ldots,x_n,\mathcal{E}\} = \int_\omega \{x_{n+1}|x_1,\ldots,x_n,\omega,\mathcal{E}\}\{\omega|x_1,\ldots,x_n,\mathcal{E}\} \quad (5.3)$$

From the conditional independence,

$$\{x_{n+1}|x_1, \ldots , x_n,\mathcal{E}\} = \int_\omega \{x_{n+1}|\omega,\mathcal{E}\}\{\omega|x_1, \ldots , x_n,\mathcal{E}\} \quad (5.4)$$

and using Bayes' Rule to evaluate $\{\omega|x_1, \ldots , x_n,\mathcal{E}\}$,

$$\{x_{n+1}|x_1,\ldots,x_n,\mathcal{E}\} = \frac{\int_\omega \{x_{n+1}|\omega,\mathcal{E}\}\{x_1,\ldots,x_n|\omega,\mathcal{E}\}\{\omega|\mathcal{E}\}}{\int_\omega \{x_1,\ldots,x_n|\omega,\mathcal{E}\}} . \quad (5.5)$$

Perfect information about exchangeable experimental trials corresponds to perfect information about ω, which specifies which long-run frequency distribution is appropriate for the sequence of exchangeable experiments. De Finetti's theorem provides a link between the subjective and frequency interpretations of probability. The concept of exchangeable experiments defines a domain in which frequency-based probabilities are appropriate. It also gives a means of interpreting the classical limit theorems of probability in a subjective context. De Finetti proves the strong law of large numbers for exchangeable random quantities ([20]). We shall use this important result extensively in the development that follows.

The meaning of de Finetti's theorem is perhaps best illustrated by the simplest case. Suppose that the exchangeable quantities x_1, x_2, ... are Bernoulli random variables, which take on only the values 1 or 0. Then de Finetti's theorem states that these random variables may be considered as independent, conditioned on a parameter

p, the probability of having a 1 result on any given trial. The prior probability distribution assigned to the result of any sequence of trials will then depend on the prior probability distribution assigned to the parameter p.

Now suppose that we observe a sequence of experimental trials $x_1, x_2, \ldots, x_n$ and build up a histogram of the results. Since there are only two possible outcomes for each trial and the number of trials is fixed, the histogram has only one degree of freedom; it is completely specified by f, the fraction of 1's. By the (strong) law of large numbers, f approaches p (with probability one) as n approaches infinity. We can think therefore of p as a constant of the process that characterizes any sequence of Bernoulli trials. By de Finetti's theorem p characterizes the sequence completely, for if we knew something beyond the value of p for a particular trial, we would destroy the exchangeability of the sequence.

For the Bernoulli case one constant p is sufficient to specify the long run frequency distribution uniquely. However, when more than two outcomes are allowed for the exchangeable experimental trials x_i, it is possible to have a set of process constants that do <u>not</u> specify the probability distribution uniquely. This concept is the key to an understanding of the applications of the maximum entropy principle.

Consider the case in which the outcome space has N points. We shall define a random variable x on these points; x can assume values $x(1), \ldots, x(N)$. The probability distribution assigned to x_j for any arbitrarily specified experimental trial j is then

$\{x_j|\mathcal{E}\} = p(x) = (p_1, \dots, p_N)$, where $p_k = \{x_j = x(k)|\mathcal{E}\}$ for $k = 1, \dots, N$.[*] The probability distribution $p_1, \dots, p_N$ will have $N-1$ degrees of freedom: the normalization constraint

$$\sum_{k=1}^{N} p_k = 1 \tag{5.6}$$

will determine one number, say p_N, in terms of the other $N-1$ probability assignments.

Let us consider a sequence of n identical, indistinguishable (i.e., exchangeable) experimental trials $x_1, \dots, x_n$. Let n_k be the number of occurrences of the k^{th} outcome, $k = 1, \dots, N$. We can define the long-run fraction

$$f_k = \frac{n_k}{n} \tag{5.7}$$

as the fraction of experiments that yield the k^{th} outcome. The histogram $f_1, \dots, f_N$ will have $N-1$ degrees of freedom as a result of the constraint

$$\sum_{k=1}^{N} n_k = n \ . \tag{5.8}$$

Each possible histogram $(f_1, \dots, f_N)$ is then contained in an $N-1$ dimensional simplex of an N-dimensional Euclidean space defined by the constraints $f_k \geq 0$, $k = 1, \dots, N$, and

$$\sum_{k=1}^{N} f_k = 1 \tag{5.9}$$

[*] Because the trials are exchangeable the distribution is the same for all j. We shall drop the subscript j henceforth.

obtained from (5.8). This simplex constitutes the outcome space for the random quantity ω.

A set of N independent equations (including (5.9)) involving the f_k would serve to specify a unique histogram. As the number of trials n approaches infinity these fractions $f_1, \ldots, f_N$ will converge to the probabilities $p_1, \ldots, p_N$ (by the law of large numbers). Therefore we can think of the probabilities $p_1, \ldots, p_N$ as specifying the process that generates the exchangeable trials. These numbers may not be known, and we may wish to infer them from data and prior information by Bayes' Rule. If we had a set of equations that we could solve for the numbers $p_1, \ldots, p_N$ such inference would be unnecessary; we could solve directly for the large sample limit of the histogram, ω. For example, knowing the first $N-1$ moments of a random variable x might allow us to solve for this limiting distribution.

5.3 The Extended Principle of Insufficient Reason as a Basis for the Maximum Entropy Principle

Suppose that a set of functions $\theta_1, \ldots, \theta_I$ are defined on the probability distribution $p_1, \ldots, p_N = \{x | \omega, \mathcal{E}\}$. Let us suppose that our prior knowledge $\mathcal{E}$ does not concern the probabilities $p_1, \ldots, p_N$ directly, but rather our knowledge relates to the values $c_1, \ldots, c_I$ attained by these functions and therefore it relates indirectly to the probabilities $p_1, \ldots, p_N$:

$$c_i = \theta_i(p_1, \ldots, p_N) = \theta_i(\{x | \omega, \mathcal{E}\}) \qquad i = 1, \ldots, I . \qquad (5.10)$$

It will be assumed that θ_i is a continuous function of the p_k's. An important special case will be that in which the functions θ_i are expectations. The expected value of a function $\varphi_i(x)$ of the random quantity x is given by

$$c_i = < \varphi_i(x) | \omega, \mathcal{E} > = \sum_{k=1}^{N} \varphi_i(x(k)) p_k \; . \qquad (5.11)$$

Suppose that the values of the constants $c_1, \ldots , c_I$ are known, but we do not know the limiting histogram ω (equivalent by the strong law of large numbers to the probability distribution $\{x | \omega, \mathcal{E}\}$). The normalization condition (5.9) plus N-1 independent ensemble constants of the form (5.10) would serve to determine the distribution ω. Obviously, there might be many alternative sets of constants that could be used to determine a given distribution. A trivial set of constants would be a set of N-1 identity equations each specifying the probability assigned to a single outcome.

We shall often be confronted with situations in which the sets of equations defining constants $c_1, \ldots , c_I$ in terms of the distribution $p_1, \ldots , p_N$ are not sufficient to specify this distribution uniquely. We may perceive our state of information to be equivalent to statements about only a few process constants $c_1, c_2, \ldots , c_I$, while the number of outcome points N is very large. It may often be desirable to let N approach infinity. Whenever N-1 is greater than I, the process constants $c_1, \ldots , c_I$ will not be sufficient to specify a unique distribution ω. In these situations we must rely on an additional invariance principle.

The invariance principle that will permit us to resolve the indeterminacy in the distribution ω for an exchangeable process is

> The Extended Principle of Insufficient Reason: Suppose that our state of knowledge about an exchangeable sequence $x_1, \ldots , x_n$ is perceived to relate to the values $c_1, \ldots , c_I$ attained by a sequence of given functions $\theta_1, \ldots , \theta_I$ defined on the long run frequency distribution ω (by the strong law of large numbers, ω is equivalent to the probability distribution $(p_1, \ldots , p_N) = \{x|\omega ,\mathcal{E}\}$ that would be assigned to a single observation on the basis of a very large number of past observations of the process). Then any two sequences of experimental outcomes whose histograms give equal values to the functions $\theta_1, \ldots , \theta_I$ are to be judged equally probable.

If the set of constants $c_1, \ldots , c_I$ is sufficient to determine the distribution ω uniquely, then the extended principle of insufficient reason reduces to the invariance inherent in an exchangeable process; the probability distribution assigned to any sequence of experimental results is invariant to permutations in the order of the sequence. That is, every sequence leading to the same histogram is equally probable. When the constants $c_1, \ldots , c_I$ do not specify a unique distribution ω through the equations (5.10), then applying the functions $\theta_1, \ldots , \theta_I$ to different histograms $f_1, \ldots , f_N$ and $f'_1, \ldots , f'_N$ resulting from two sequences of experimental outcomes may lead to the same values for these functions: $\theta_i(f_1, \ldots , f_N) = \theta_i(f'_1, \ldots , f'_N)$. These two sequences of experimental outcomes are then to be judged equally probable.

The extended principle of insufficient reason is an invariance principle that may be derived from the basic desideratum. The basis for this derivation is the way in which we have specified our state of information. We are assuming that prior knowledge $\mathcal{E}$ relates only to the values $c_1, \ldots , c_I$ attained by the functions $\theta_1, \ldots , \theta_I$ when applied to the limiting histogram ω that summarizes a large number of experimental outcomes. Therefore, our state of information should be unchanged by any transformation on experimental outcomes that leaves the values of the functions θ_i unchanged. Suppose we have an arbitrary sequence $x_1^o, \ldots , x_n^o$ of outcomes generated by the exchangeable process, and we construct a histogram $f_1, \ldots , f_N$ summarizing the fractional number of times the k^{th} outcome was observed, $k = 1, \ldots , N$. Now suppose that a transformation or relabeling of the outcomes leads to a new histogram $f_1', \ldots , f_N'$ that gives the same values to the functions $\theta_1, \ldots , \theta_I$ (the transformation depends on the functions $\theta_1, \ldots , \theta_I$, but of course it cannot depend on the particular sequence $x_1^o, \ldots , x_n^o$). Since prior information relates only to the values attained by the functions $\theta_1, \ldots , \theta_I$, the transformation leaves the state of information unchanged. Hence we must assign the same probability in both situations as a consequence of the basic desideratum: We must assign the same probability to the transformed sequence as to the original sequence.

A limitation of the specified state of knowledge to a subset of the constants needed to specify the distribution may be regarded as a "null hypothesis." If the distribution that results from applying the extended principle of insufficient reason is confirmed by the data, then we have a relatively simple characterization of an apparently complex process.* If the distribution does not correspond to the observed data, then we must try to find a better model, perhaps by re-examining the way we have specified our knowledge and incorporating aspects of our information that were previously ignored. We shall return to this question of model testing in Chapter 7.

From the extended principle of insufficient reason it is a straightforward matter to derive the maximum entropy principle. In fact, the mathematics of the proof are standard in physics ([31],[77]).

Theorem 5.1: The Maximum Entropy Principle for Statistical Ensembles.

Suppose $c_1, \dots, c_I$ are known constants corresponding to functions $\theta_1, \dots, \theta_I$ acting on the (unknown) distribution ω. The extended principle of insufficient reason implies a maximum entropy principle for statistical ensembles. Namely, the distribution $\{x | c_1, \dots, c_I = \mathcal{E}\} = (p_1, \dots, p_N)$ that should be assigned to x on the basis of knowledge of the constants $c_1, \dots, c_I$ is that which maximizes

$$- \sum_{k=1}^{N} p_k \log p_k \qquad (5.12)$$

* A similar viewpoint on the maximum entropy principle (to be derived from the extended principle of insufficient reason) has been expressed by Good [28] and Jaynes [35], [40]. It is especially important in physics where knowledge concerns conserved quantities such as energy and angular momentums.

subject to

$$p_k \geq 0 \qquad k = 1, \ldots , N \tag{5.13}$$

$$\sum_{k=1}^{N} p_k = 1 \tag{5.14}$$

and

$$\theta_i(p_1, \ldots , p_N) = c_i \qquad i = 1, \ldots , I . \tag{5.15}$$

<u>Proof of Theorem 5.1</u>: Let us consider the situation in which we have observed n exchangeable trials, i.e., we have n samples from the stochastic process $x_1, x_2, \ldots$. The k^{th} outcome has been observed to occur n_k times, $k = 1, \ldots , N$. The histogram corresponding to these results might have been generated by any of W sequences, where

$$W = \frac{n!}{n_1! \; n_2! \cdots n_N!} \tag{5.16}$$

is the number of ways of arranging n objects in N categories, with n_1 objects in the first category, n_2 in the second, etc. Because the trials are exchangeable, each of these sequences is equally probable.

Suppose that n, n_k for $k = 1, \ldots , N$ is large compared to N, i.e., $n >> N$. Then the factorials in (5.16) may be evaluated by Stirling's approximation:

$$n! \approx e^{-n} n^n \sqrt{2\pi n} \tag{5.17}$$

then

$$\log n! = n \log n - n + O(\log n)$$

where $O(\log n)$ denotes terms that increase no faster than as $\log n$ as n becomes large.

We can then write the logarithm of (5.16) as

$$\begin{aligned}\log W &= \log(n!) - \sum_{k=1}^{N} \log(n_k!) \\ &= n \log n - n - \sum_{k=1}^{N} n_k \log n_k + \sum_{k=1}^{N} n_k + O(\log n) \\ &= - \sum_{k=1}^{N} n_k \log(n_k/n) + O(\log n) \qquad (5.18)\end{aligned}$$

and dividing by the number of trials n,

$$\begin{aligned}\frac{1}{n} \log W &= - \sum_{k=1}^{N} \left(\frac{n_k}{n}\right) \log\left(\frac{n_k}{n}\right) + O(\frac{\log n}{n}) \\ &= - \sum_{k=1}^{N} f_k \log f_k + O(\frac{\log n}{n}) \ .\end{aligned}$$

As n goes to infinity, the terms of the order of $\log n/n$ become negligible compared to the first term, and by the (strong) law of large numbers the frequencies f_k converge to the probabilities p_k (with probability one). Hence

$$\begin{aligned}\lim_{n \to \infty} (\frac{1}{n} \log W) &= - \sum_{k=1}^{N} p_k \log p_k \\ &= H(p_1, \dots , p_N) \ . \end{aligned} \qquad (5.19)$$

The entropy function $H(p_1, \dots , p_N)$ measures the limit of the natural logarithm of the number of possible sequences generating a given (long run frequency) distribution divided by the sequence length.

Now let us consider the constraints imposed by knowledge of process constants, $c_1, \dots, c_I$. For a sequence of n experimental trials, we can compute corresponding experimental values $c_i(n)$ from the functions θ_i applied to the histogram $f_1, \dots, f_N$:

$$\theta_i(f_1, \dots, f_N) = c_i(n) . \tag{5.20}$$

Suppose we choose small intervals $[c_i - \delta c_i, c_i + \delta c_i]$ around the known values c_i of the i^{th} process constant, $i = 1, \dots, I$. We can choose n large enough so that each $c_i(n)$ is contained in the corresponding interval with probability $1-\epsilon$, where ϵ is an arbitrarily assigned small number; this follows from the strong law of large numbers and the assumption that the θ_i are continuous.

$$c_i - \delta c_i \leq \theta(f_1, \dots, f_N) \leq c_i + \delta c_i . \tag{5.21}$$

Now, having chosen n, let us compute the most probable histogram. By the extended principle of insufficient reason, all sequences that satisfy the given constraints (5.21) are equally probable; however, a given histogram $f_1, \dots, f_N$ may be generated by any one of $W(f_1, \dots, f_N)$ sequences. If we then search for the histogram which can be realized in the greatest number of ways consistent with the constraints (5.21), we have the following problem:

$$\text{maximize} \quad W(f_1, \dots, f_N)$$

subject to (5.21) and the conditions necessary to insure that $f_1, \dots, f_N$ is a histogram:

$$f_k \geq 0, \quad k = 1, \ldots , N, \quad \sum_{k=1}^{N} f_k = 1 . \tag{5.22}$$

For fixed n, maximizing

$$H(f_1, \ldots , f_N) = \frac{1}{n} \log W(f_1, \ldots , f_N) \tag{5.23}$$

will give the same distribution as maximizing W. If we let n go to infinity, the probability that a constraint equation (5.21) will be violated approaches zero, and the frequencies f_k converge to the probabilities p_k. We can shrink the intervals 2δ in the constraint equation to zero, and (5.23), (5.22), and (5.21) become (5.12), (5.13), (5.14), and (5.15).

Recall that W is the number of outcome sequences giving rise to a given histogram. Then dW/W would be the percentage change in the number of sequences corresponding to a given histogram. But since

$$W(f_1, \ldots , f_N) = e^{nH(f_1,\ldots,f_N)} , \tag{5.24}$$

then

$$\frac{dW}{W} = ndH . \tag{5.25}$$

As n goes to infinity, the percentage change in W as a function of changes in the histogram becomes infinitely sharp; the histogram corresponding to the distribution that solves (5.12)-(5.15) can be produced by infinitely more sequences than any histogram found by varying that distribution a finite amount.

The sharpness of the maximum in W is another manifestation of the law of large numbers. It is because of the law of large numbers that the extended principle of insufficient reason leads to the precise characterization of the distribution $\{x|c_1, \ldots , c_I = \mathcal{E}\}$ by means of the maximum entropy principle.

5.4 Solutions for Expectation Information

Let us first consider a simple example. Suppose we have an ensemble of random variables $x_1, x_2, \ldots$ that can take on N discrete real values from zero to some large positive number. Suppose only one ensemble constant is assumed, namely, the mean or ensemble expectation m. The corresponding distribution $\{x|m,\mathcal{E}\} = p(x)$ for this case is easily found using the maximum entropy principle that we have just derived from the extended principle of insufficient reason. We wish to maximize

$$H(\{x|m,\mathcal{E}\}) = -\sum_x p(x) \log p(x) \qquad (5.26)$$

subject to

$$p(x) \geq 0 \quad \text{for all outcome points} \quad x \qquad (5.27)$$

$$\sum_x p(x) = 1 \qquad (5.28)$$

and the equation defining the ensemble constant m

$$\sum_x xp(x) = m \ . \qquad (5.29)$$

We shall find that (5.27) is automatically satisfied because of the form of the entropy functional H. To deal with the other constraints we form the Lagrangian

$$\mathcal{L}(p(x),\mu,\lambda) = -\sum_x p(x)\ \log p(x) + \mu\left(\sum_x p(x)-1\right) + \lambda\left(\sum_x xp(x)-m\right) \ . \tag{5.30}$$

Setting the partial derivative of $\mathcal{L}$ with respect to $p(x)$ equal to zero gives the equation for an extremum:

$$\log p(x) + 1 - \mu - \lambda x = 0 \quad \text{for all outcome points} \quad x \ . \tag{5.31}$$

By examining the second partial derivatives of $\mathcal{L}$ we may verify immediately that this extremum will be a maximum. The Lagrange multiplier μ corresponds to the normalization condition (5.28), and so we may obtain from (5.31)

$$p(x) = \frac{1}{Z}\, e^{+\lambda x}$$

where the "partition" function $Z = Z(\lambda)$ is defined so as to normalize the probability distribution:

$$Z(\lambda) = \sum_x e^{+\lambda x} \ .$$

The multiplier λ may be evaluated in terms of the known quantity m by using (5.29):

$$\frac{\partial}{\partial\lambda} \log Z(\lambda) = \frac{1}{Z}\sum_x xe^{\lambda x} = < x|m,\mathcal{E}> = m \ . \tag{5.32}$$

In order to continue we need a specific form for $Z(\lambda)$. Let us assume that the possible outcomes x are spaced evenly and very close together, and we shall let N go to infinity. Then the sums over x

can be replaced by definite integrals over x from zero to infinity.*
We have

$$Z(\lambda) = \int_0^\infty e^{\lambda x} dx = -\frac{1}{\lambda} \qquad (5.33)$$

and

$$\frac{\partial}{\partial\lambda} \log Z(\lambda) = \frac{-1}{\lambda} = m \ . \qquad (5.34)$$

The distribution $\{x|m,\mathcal{E}\} = \frac{1}{m} e^{-x/m}$ is now clearly recognizable as the exponential distribution. For the distribution of particle energies in statistical mechanics it is called the Boltzmann distribution, and in Gibbs' formulation it is equivalent to the canonical distribution. This distribution forms the basis for most of classical statistical mechanics. With some minor adjustments having to do with the character of the outcome space the method works equally well in quantum statistical mechanics ([82],[84],[87]).

Many other well-known probability distributions can be derived on the basis of maximizing entropy subject to a knowledge of a few ensemble constants, corresponding to expectations of simple functions. Some of these distributions are listed in table 5.1:

* This artifice of allowing the outcome space to become continuous only at the stage of calculating the partition function Z avoids some problems that arise in defining the entropy functional for a continuous outcome space. An introduction to some of the literature on these problems is found in Abramson [1], pp. 39-40. A way of solving them by incorporating a measure function into the entropy functional has been proposed by Jaynes [36], [40]. For most applications of interest this measure (density of outcome points per unit interval) will be constant and the solution of the maximum entropy problem for an ensemble will be formally the same for a continuous distribution as for the discrete case.

Table 5.1

Distributions Derivable from Maximizing Entropy Subject to Knowledge of Ensemble Expectations

Outcome Space	Known Ensemble Expectations of Functions $\varphi_i(x)$	Distribution
finite interval	none	uniform
non-negative part of real line	$\varphi_1(x) = x$	exponential
real line	$\varphi_1(x) = x$ $\varphi_2(x) = x^2$	normal
real line	$\varphi_1(x) = x$ $\varphi_2(x) = \log x$	gamma
[0,1]	$\varphi_1(x) = \log x$ $\varphi_2(x) = \log (1-x)$	beta
real line	$\varphi_1(x) = \log x$ $\varphi_2(x) = x^\beta$ (β is a fixed constant)	Weibull
real line	$\varphi_1(x) = \lvert x\rvert$	Laplace

It is possible to show in general that maximizing entropy subject to expectation constraints results in probability distributions of a special exponential form. Moreover, any distribution of this form can be regarded as the solution to a problem of maximizing entropy subject to expectation constraints.

Theorem 5.2: Given an ensemble (a process generating exchangeable trials) $x_1, x_2, \ldots,$ then for the distribution $\{x|\mathcal{E}\} = p(x)$ to be of the exponential form

$$p(x) = \frac{1}{Z(\lambda_1, \dots, \lambda_I)} \exp\left(\sum_{i=1}^{I} \lambda_i \varphi_i(x) \right) \qquad (5.35)$$

it is necessary and sufficient that the distribution $p(x)$ maximize the entropy functional

$$H(p(x)) = - \int_x p(x) \log p(x) \qquad (5.36)$$

subject to constraints on the expected values of I functions φ_i, $i = 1, \dots, I$, of the random variable x:

$$\int_x \varphi_i(x) p(x) = < \varphi_i(x) | \mathcal{E} > = c_i \ . \qquad (5.37)$$

It is assumed that these expected values remain finite if we allow outcome spaces to become infinite.

Sufficiency

Suppose we have the following optimization problem: Maximize

$$H(p(x)) = - \int_x p(x) \log p(x) \qquad (5.38)$$

subject to

$$p(x) \geq 0 \quad \text{all outcome points} \quad x \qquad (5.39)$$

$$\int_x p(x) = 1 \qquad (5.40)$$

$$\int_x \varphi_i(x) p(x) = c_i \qquad i = 1, \dots, I \ . \qquad (5.41)$$

In the usual way we form the Lagrangian:

$$\mathcal{L}(p(x), \mu, \lambda_1, \dots , \lambda_I) = - \int_x p(x) \log p(x) + \mu \left(\int_x p(x) - 1 \right) \tag{5.42}$$
$$+ \sum_{i=1}^{I} \lambda_i \left(\int_x p(x) \varphi_i(x) - c_i \right) .$$

The condition for a maximum is easily seen to be that

$$\frac{\partial \mathcal{L}}{\partial p(x)} = 0 = -\log p(x) - 1 + \mu + \sum_{i=1}^{I} \lambda_i \varphi_i(x) \tag{5.43}$$

hold at every outcome point x. This is equivalent to the form (5.35), where the constant $Z(\lambda_1, \dots , \lambda_I)$ is determined by the normalization condition (5.40). The values of the multipliers $\lambda_1, \dots , \lambda_I$ may be solved in terms of the expectation values c_i using the relations that

$$\frac{\partial}{\partial \lambda_i} \log Z(\lambda_1, \dots , \lambda_I) = < \varphi_i(x) | \mathcal{E} > = c_i . \tag{5.44}$$

It is presumed that at least one distribution satisfies the constraint equations (5.39), (5.40), (5.41). Since the functional H is strictly concave in the $p(x)$, the solution to the maximization problem will be uniquely determined by (5.43) and the normalization condition (5.40). The system of equations (5.44) must therefore determine a unique distribution.

<u>Necessity</u>:

Suppose we are given a distribution of the form (5.35), say, $p'(x)$. We can compute the expectations of the functions φ_i over this distribution:

$$< \varphi_i(x) | \mathcal{E} > = c_i . \tag{5.45}$$

These are presumed by hypothesis to be finite. Now supposing we solve the optimization problem specified by (5.38), (5.39), (5.40), (5.41): at least one distribution, namely $p'(x)$, satisfies the constraints (5.41), so by the sufficiency part of the proof we obtain an answer

$$p*(x) = \frac{1}{Z*(\lambda_1^*, \dots, \lambda_n^*)} \exp\left(\sum_{i=1}^{I} \lambda_i^* \varphi_i(x) \right) . \qquad (5.46)$$

Z and $Z*$ are both determined by the requirement (5.40) that the distribution must normalize to one. For both distributions we must have

$$< \varphi_i(x) | \mathcal{E} > = c_i = \frac{\partial}{\partial \lambda_i} \log Z \qquad (5.47)$$

which must determine the distribution uniquely because the maximization problem has a unique solution. Hence $p(x)$ and $p*(x)$ must be identical.

An alternative argument (Jaynes [36]) may give more insight into the theorem. Let us start with the fact that $\log z \geq (1-z^{-1})$, with equality if and only if $z = 1$. Then if we have any two probability distributions $p(x)$, $p*(x)$ over an outcome space,

$$\int_x p(x) \log\left|\frac{p(x)}{p*(x)}\right| \geq \int_x p(x)\left|1 - \frac{p*(x)}{p(x)}\right| = 0$$

or $\qquad (5.48)$

$$-\int_x p(x) \log p(x) \leq -\int_x p(x) \log p*(x)$$

with equality if and only if $p(x) = p*(x)$. This result is equivalent to lemma 3 of Section 4.

Now suppose we choose

$$p^*(x) = \frac{1}{Z(\lambda_1, \dots, \lambda_I)} \exp\left(\sum_{i=1}^{I} \lambda_i \varphi_i(x) \right) \qquad (5.49)$$

where $\lambda_1, \dots, \lambda_I$ are fixed constants, and $Z(\lambda_1, \dots, \lambda_I)$ is determined by the normalization condition:

$$Z(\lambda_1, \dots, \lambda_I) = \int_X \exp\left(\sum_{i=1}^{I} \lambda_i \varphi_i(x) \right) . \qquad (5.50)$$

Inserting $p^*(x)$ in (5.48) we obtain

$$- \int_X p(x) \log p(x) \leq \int_X p(x) \left[\sum_{i=1}^{I} \lambda_i \varphi_i(x) - \log Z(\lambda_1, \dots, \lambda_I) \right] . \qquad (5.51)$$

Let us require that $p(x)$ satisfy the constraints

$$\int_X p(x) \varphi_i(x) = c_i \qquad i = 1, \dots, I . \qquad (5.52)$$

Equation (5.51) then yields

$$H(p(x)) \leq \sum_{i=1}^{I} \lambda_i c_i - \log Z(\lambda_1, \dots, \lambda_I) . \qquad (5.53)$$

The maximum value of the entropy H is then attained if and only if $p(x) = p^*(x)$, in which case (5.53) becomes an equality. The only remaining detail is to choose the constants $\lambda_1, \dots, \lambda_I$ so that (5.52) is satisfied; this condition implies

$$c_i = \frac{\partial}{\partial \lambda_i} \log Z(\lambda_1, \dots, \lambda_I) . \qquad (5.54)$$

The equations (5.54) must determine the λ_i so that the resulting distribution $p*(x)$ is unique. For if there were two distributions of the form (5.49) such that the fixed constants λ_i satisfied (5.54), we could choose $p*(x)$ equal to one of them arbitrarily and (5.53) would imply that this distribution had greater entropy than the other. Since the choice of the distribution is arbitrary we have a contradiction, and the distribution $p*(x)$ must be unique.

The implications of theorem 5.2 become apparent when we consider the situation in which some of the constants c_i are not known. That is, we know the number of such ensemble constants, and for each we know the corresponding function φ_i, but we do not know the value of the constant c_i. We can think of the c_i (or equivalently, the λ_i determined from equation (5.54)) as unknown parameters of a probability distribution whose functional form is otherwise known (e.g., the functional form is derived from the extended principle of insufficient reason by means of the maximum entropy principle).

The usual precepts of the Bayesian methodology imply encoding knowledge about unknown parameters $\alpha_1, \ldots, \alpha_\nu$ in the form of a probability distribution $\{\alpha_1, \ldots, \alpha_\nu|\mathcal{E}\}$, then updating this distribution using Bayes' rule as samples $x_1, x_2, \ldots$ from the ensemble become available. If the distribution $\{x|\alpha_1, \ldots, \alpha_\nu, \mathcal{E}\}$ is characterized by sufficient statistics, the effect of the samples in the updating process can be summarized by a vector of fixed length and it will be possible to construct conjugate families of probability distributions for the parameters.

Suppose we have samples $x_1, \ldots, x_n$ from an ensemble, and suppose there exist m functions of the observations, $s_k(x_1, \ldots, x_n)$,

$k = 1, \dots, m$. We shall suppose that m is independent of the number of samples n. Since the parameters may be varied independently we shall wish to have $m \geq \nu$, the dimension of the parameter space. The functions s_k are defined to be sufficient statistics if they include all the necessary information for updating the prior distribution on the parameters $\alpha_1, \dots, \alpha_\nu$. That is, if

$$s_k(x_1, \dots, x_n) = s_k(x^*_1, \dots, x^*_n) \qquad (5.55)$$

for $k = 1, \dots, m$, then the ratio of the likelihood functions for any two sets of parameter values is the same for $x_1, \dots, x_n$ as for $x^*_1, \dots, x^*_n$.

$$\frac{\prod_{i=1}^{n} \{x_i|\alpha_1, \dots, \alpha_\nu, \mathcal{E}\}}{\prod_{i=1}^{n} \{x_i|\alpha'_1, \dots, \alpha'_\nu, \mathcal{E}\}} = \frac{\prod_{i=1}^{n} \{x^*_i|\alpha_1, \dots, \alpha_\nu, \mathcal{E}\}}{\prod_{i=1}^{n} \{x^*_i|\alpha'_1, \dots, \alpha'_\nu, \mathcal{E}\}} \qquad (5.56)$$

where $\alpha_1, \dots, \alpha_\nu$ and $\alpha'_1, \dots, \alpha'_\nu$ are any two possible (i.e., non-zero probability) sets of parameter values. From this definition it is apparent that the sufficient statistics $s_1, \dots, s_m$ partition the possible sets of observations $x_1, \dots, x_n$ into equivalence classes that give the same ratio for the likelihood functions associated with any two parameter values $\alpha_1, \dots, \alpha_\nu$ and $\alpha'_1, \dots, \alpha'_\nu$. It is convenient to include the number of observations n as an ancillary statistic, as the likelihood ratio (5.56) will in general depend on this quantity.

There is an important theorem that gives necessary and sufficient conditions for the existence of sufficient statistics. This theorem was proved almost simultaneously by B. O. Koopman [46] and E. J. G. Pitman [63], but the full power of the theorem only became evident with the proof by Jeffreys for discrete as well as continuous outcome spaces [42], [43]. The proof given here is substantially the same as Jeffreys' proof.

Theorem 5.3 (Koopman-Pitman): Suppose that a probability distribution has parameters $\vec{\alpha} = \alpha_1, \ldots, \alpha_\nu$. A necessary and sufficient condition that this probability distribution possess sufficient statistics for inference on these parameters is that it be expressible in the form

$$\{x|\vec{\alpha},\mathcal{E}\} = A(\vec{\alpha})\psi(x)\exp\left(\sum_{i=1}^{\nu} u_i(\vec{\alpha})v_i(x)\right) . \tag{5.57}$$

Proof: Suppose that the probability distribution has the form (5.57). Consider a set of n independent observations from this distribution, $x_1, \ldots, x_n$. Clearly the likelihood function for these observations is

$$\begin{aligned}\{x_1, \ldots, x_n|\vec{\alpha},\mathcal{E}\} &= \prod_{j=1}^{n} \{x_j|\vec{\alpha},\mathcal{E}\} \\ &= A^n(\vec{\alpha})\left(\prod_{j=1}^{n} \psi(x_j)\right) \exp\left(\sum_{i=1}^{\nu} u_i(\vec{\alpha}) \sum_{j=1}^{n} v_i(x_j)\right) .\end{aligned} \tag{5.58}$$

Since in the ratio of likelihood functions for different sets of parameter values $\vec{\alpha}, \vec{\alpha}'$ the common factor $\prod_{j=1}^{n} \psi(x_j)$ will cancel, the ν functions

$$s_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} v_i(x_j) \qquad i = 1, \ldots, \nu \tag{5.59}$$

are seen to be sufficient statistics for the parameters $\alpha_1, \ldots, \alpha_\nu$.

The necessity portion of the proof requires a considerably more involved argument. Let us suppose that we are given a set of m functions $s_k(x_1, \ldots, x_n)$, $k = 1, \ldots, m$, such if $\vec{x} = x_1, \ldots, x_n$ and $\vec{x}* = x_1^*, \ldots, x_n^*$ are any two sets of possible observations for which $s_k(\vec{x}) = s_k(\vec{x}*)$, $k = 1, \ldots, m$, then for any two possible parameter sets $\vec{\alpha}$ and $\vec{\alpha}'$, equation (5.56) is satisfied. We shall assume that the number of observations n is greater than the number of statistics m.

Let us denote the likelihood function for the observations by L. Since the observations are independent,

$$L = \{x_1, \ldots, x_n|\vec{\alpha},\mathcal{E}\} = \prod_{j=1}^{n} \{x_j|\vec{\alpha},\mathcal{E}\} \qquad (5.60)$$

and therefore its logarithm must be of the form

$$\log L = \sum_{j=1}^{n} g(x_j, \alpha_1, \ldots, \alpha_\nu) \qquad (5.61)$$

where g is a known function of the $\nu+1$ arguments.

The knowledge that the sufficient statistics $s_1, \ldots, s_m$ exist permits us to write another expression for $\log L$. Since the ratios in (5.56) can depend only on the values of the statistics $s_1, \ldots, s_m$, L must be a product of two factors, one depending on $\vec{\alpha}$ and the observations only through the statistics $s_1, \ldots, s_m$, and the second whose dependence on the observations $x_1, \ldots, x_n$ is unrestricted, but which cannot depend on the parameters $\vec{\alpha}$; this second factor must cancel in computing the likelihood ratios in (5.56). Therefore, the logarithm of the likelihood function must be expressible in the form

$$\log L = \Phi(s_1, \ldots, s_m, \alpha_1, \ldots, \alpha_\nu) + \chi(x_1, \ldots, x_n) \; . \qquad (5.62)$$

The required equivalence of the functional forms (5.61) and (5.62) implies that we may use an equivalent set of sufficient statistics that are additive in form. Suppose that we start with a fixed set of parameter values $\vec{\alpha}*$, $\vec{\alpha}* = \alpha_1^*, \ldots, \alpha_\nu^*$, and we change one parameter at a time to another possible parameter value, $\alpha_i^* + \delta_i$. These changes need not be infinitesimal. For a given set of observations $x_1, \ldots, x_n$ the corresponding change in $\log L$ is (from (5.61))

$$\begin{aligned}\Delta_i &= \log L(x_1,\ldots,x_n,\alpha_1^*,\ldots,\alpha_i^* + \delta_i,\ldots,\alpha_\nu^*) \\ &\qquad - \log L(x_1,\ldots,x_n,\alpha_1^*,\ldots,\alpha_i^*,\ldots,\alpha_\nu^*) \\ &= \sum_{j=1}^{n} \Big(g(x_j,\alpha_1^*,\ldots,\alpha_i^* + \delta_i,\ldots,\alpha_\nu^*) - g(x_j,\alpha_1^*,\ldots,\alpha_i^*,\ldots,\alpha_\nu^*) \Big) \qquad (5.63) \\ &\triangleq \sum_{j=1}^{n} v_i(x_j) \; .\end{aligned}$$

The functions $v_i(x_j)$ defined by (5.63) are clearly dependent on the values chosen for $\vec{\alpha}*$ and δ_i. But we observe that Δ_i is really the logarithm of the likelihood ratio for a special case of (5.56). We can see the consequences of this fact even more clearly from (5.62): we require that

$$\Delta_i = \Phi(s_1,\ldots,s_m,\alpha_1^*,\ldots,\alpha_i^* + \delta_i,\ldots,\alpha_\nu^*) - \Phi(s_1,\ldots,s_m,\alpha_1^*,\ldots,\alpha_i^*,\ldots,\alpha_\nu^*) \; . \qquad (5.64)$$

In words, the Δ_i, which are functions of the observations of the form (5.63) must depend on the observations in the same ways as the statistics $s_1, \ldots, s_m$. Inserting the functions $s_k(x_1, \ldots, x_n)$, $k = 1, \ldots, m$ in (5.64) must reduce the equations (5.64) to an identity in the observations $x_1, \ldots, x_n$ that holds regardless of the values chosen for $\vec{\alpha}*$ and $\delta_1, \ldots, \delta_\nu$. Therefore, the functions $s_1, \ldots, s_m$ and $\Delta_1, \ldots, \Delta_\nu$ must determine the same equivalence classes on the observations, regardless of the values chosen for $\vec{\alpha}*$ and $\delta_1, \ldots, \delta_\nu$. We have now found a set of ν functions $\Delta_i(x_1, \ldots, x_n)$ that satisfy the definition of sufficient statistics, and have the additive form (5.63). For the remainder of the proof we shall use the sufficient statistics Δ_i, $i = 1, \ldots, \nu$, defined by (5.63).

Let us now consider two possible sets of parameter values, $\vec{\alpha}$ and $\vec{\alpha}*$. The logarithm of the likelihood ratio for a specific set of observations $\vec{x} = x_1, \ldots, x_n$ must depend on these observations only through the sufficient statistics Δ_i:

$$\begin{aligned}\log L(\vec{x},\vec{\alpha}) - \log L(\vec{x},\vec{\alpha}*) &= \sum_{j=1}^{n} g(x_j,\vec{\alpha}) - g(x_j,\vec{\alpha}*) \\ &= \Phi(\Delta_1, \ldots, \Delta_\nu,\vec{\alpha}) - \Phi(\Delta_1, \ldots, \Delta_\nu,\vec{\alpha}*)\end{aligned} \tag{5.65}$$

since the Δ_i are sufficient statistics. From (5.63),

$$\begin{aligned}\log L(\vec{x},\vec{\alpha}) - \log L(\vec{x},\vec{\alpha}*) = \Phi\left(\sum_{j=1}^{n} v_1(x_j), \ldots, \sum_{j=1}^{n} v_\nu(x_j),\vec{\alpha}\right) \\ - \Phi\left(\sum_{j=1}^{n} v_1(x_j), \ldots, \sum_{j=1}^{n} v_\nu(x_j),\vec{\alpha}*\right)\end{aligned} \tag{5.66}$$

$$\triangleq \Psi\left(\sum_{j=1}^{n} v_1(x_j), \ldots, \sum_{j=1}^{n} v_\nu(x_j),\vec{\alpha}*\right) \tag{5.67}$$

taking the parameter values $\vec{\alpha}*$ as a fixed reference point.

Suppose now that the j^{th} observation x_j is replaced by $x_j + h_j$. The change in the logarithm of the likelihood ratio must be the same whether computed from (5.67) or from the form (5.61). Defining a difference operator D_j on functions of $x_1, \ldots , x_n$ by

$$D_j f(x_1,\ldots,x_n) = f(x_1,\ldots,x_j + h_j,\ldots,x_n) - f(x_1,\ldots,x_j,\ldots,x_n) \quad (5.68)$$

we can write this equivalence as

$$\begin{aligned} &D_j g(x_j,\vec{\alpha}) - D_j g(x_j,\vec{\alpha}*) \\ &\qquad = D_j \Psi\Big(\sum_{k=1}^{n} v_1(x_k), \ \ldots \ , \sum_{k=1}^{n} v_\nu(x_k)\Big) \ . \end{aligned} \quad (5.69)$$

We note that the left-hand side depends only on the one observation x_j and not on any other observations. Hence the right-hand side depends on the observations through the functions $\Delta_i = \sum\limits_{k=1}^{n} v_i(x_k)$, $i = 1, \ldots , \nu$. If we were to vary another observation $x_{j'}$, the right-hand side of (5.69) must remain the same, so the change in the functions $\Delta_i = \sum\limits_{k=1}^{n} v_i(x_k)$ must cancel out. This cancellation will occur if and only if Ψ is linear in the functions Δ_i. (Otherwise a variation in x_j would produce equal changes in the arguments of the two functions Ψ whose difference is taken by the operator D_j. There is no functional dependence between the Δ_i, and therefore the right-hand side of (5.67) should not be the same.) Then

$$\Psi(\Delta_1, \ \ldots \ , \Delta_\nu,\vec{\alpha}) = \sum_{i=1}^{\nu} u_i(\vec{\alpha})\Delta_i + p(\vec{\alpha}) \ , \quad (5.70)$$

with

$$\Delta_i = \sum_{j=1}^{n} v_i(x_j) \tag{5.71}$$

and since Ψ is in fact the logarithm of the likelihood function:

$$\Psi = \Phi(\Delta_1, \ldots, \Delta_\nu, \vec{\alpha}) - \Phi(\Delta_1, \ldots, \Delta_\nu, \vec{\alpha}*) \tag{5.72}$$

$$= \sum_{j=1}^{n} g(x_j, \alpha_1, \ldots, \alpha_\nu) - \sum_{j=1}^{n} g(x_j, \alpha_1^*, \ldots, \alpha_\nu^*) \tag{5.73}$$

and so we must have

$$\begin{aligned} g(x_j, \alpha_1, \ldots, \alpha_\nu) = \sum_{i=1}^{\nu} u_i(\vec{\alpha}) v_i(x_j) + p(\vec{\alpha}) \\ + g(x_j, \alpha_1^*, \ldots, \alpha_\nu^*) \ . \end{aligned} \tag{5.74}$$

The parameter set $\alpha_1^*, \ldots, \alpha_\nu^*$ is fixed, and since $g(x_j, \alpha_1, \ldots, \alpha_\nu)$ is the logarithm of the likelihood function for a single observation, taking its exponential gives us the desired form (5.57).

There is one difficulty in the derivation above that must be clarified before we can regard the theorem as proved. Some parameter values α_i may determine regions in which the observations $x_1, \ldots, x_n$ will be impossible, e.g., have probability zero. The likelihood functions corresponding to different sets of parameter values but the same observations might have a ratio that is zero or infinite. The logarithm of this ratio will be indeterminate, and the derivation that we have just presented will fail to hold.

The difficulty may be remedied quite easily if we can separate those parameters that define regions of zero probability. Consider the case in which

$$\{x_j|\vec{\alpha},\mathcal{E}\} = 0 \quad \text{if} \quad x \geq \beta(\vec{\alpha}) \ . \tag{5.75}$$

Let us suppose that it is possible to transform to a new parametrization of the likelihood function in which β is a parameter, say $\beta = \alpha_1$. Then the likelihood function for n observations is of the form

$$\{x_1, \dots , x_n|\vec{\alpha},\mathcal{E}\} = \begin{cases} \prod_{j=1} f(x_j,\vec{\alpha}) & \text{if} \quad x_j \geq \alpha_1 \\ & \quad \text{for all} \quad j = 1, \dots , n \\ 0 & \text{if} \quad \min\limits_{j=1,\dots,n} x_j < \alpha_1 \end{cases} \tag{5.76}$$

where $f(x_j,\vec{\alpha})$ is a known function of the j^{th} observation x_j and the parameters $\alpha_1, \dots , \alpha_\nu$.*

From the definition given for sufficient statistics it is clear that any set of sufficient statistics must include a function of the form

$$s_1(x_1, \dots , x_n) = \min_{j=1,\dots,n} x_j \ . \tag{5.77}$$

For if in (5.75) we vary the parameter α_1 for a fixed set of observations $x_1, \dots , x_n$, the likelihood function goes to zero

* The assumption that we can parametrize the likelihood function in the form (5.76) can be regarded as part of the hypothesis of the theorem. An extension of the Koopman-Pitman theorem to arbitrary parametrizations of boundaries of the outcome space may be possible, but the version we have proved here covers all of the probability distributions in general use.

discontinuously as α_1 exceeds $\min x_j$. Two sets of observations that differ on $\min x_j$ cannot therefore be in the same equivalence class; they would lead to different results for the updating of a probability distribution on the parameters $\alpha_1, \ldots , \alpha_\nu$.

Consider the likelihood ratio for any two sets of parameter values, as in (5.56), for which the sufficient statistic $\min x_j$ satisfies

$$\min_j x_j \geq \max(\alpha_1, \alpha_1') \ . \tag{5.78}$$

Providing we consider only the parameter values and observations such that (5.78) is satisfied, the derivation given previously shows that the existence of sufficient statistics implies the exponential form (5.57).*

The extension to additional constraints of the form (5.75) is straightforward; if the inequality were reversed we would have as a sufficient statistic the maximum observation rather than the minimum. In the most general case for a one-dimensional observation x_j the outcome space would be contained in a sequence of disjoint intervals whose endpoints were given by parameters $\alpha_1, \ldots , \alpha_k$.

In the course of our proof we have established the following important corollary to the Koopman-Pitman theorem:

Corollary 5.3.1: Given that sufficient statistics exist for a probability distribution $\{x|\alpha_1, \ldots , \alpha_\nu, \mathcal{E}\}$ whose region of zero probability is independent of the parameter values, then a new set of ν

* The derivation is exactly as previously stated except that α_1 and the observation generating the minimum are held fixed; (by virtue of exchangeability we can take this minimum observation to be x_1, so we are setting $\delta_1 = 0$ and $h_1 = 0$). For details see Jeffreys [42].

sufficient statistics may be written in the form

$$s_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} v_i(x_j) \qquad i = 1, \ldots, \nu . \qquad (5.79)$$

Let us now examine the implications of the Koopman-Pitman theorem together with theorem 5.2, the solution for the maximum entropy distribution subject to expectation constraints. We find an equivalence between the forms (5.56) and (5.35). Any distribution of the form (5.57) clearly implies the form (5.35), and any distribution of the form (5.57) can be reduced to the form (5.35) by inducing a new function $\varphi_{I+1}(x) = \log \psi(x)$, and bringing this quantity into the exponent. This procedure may involve the addition of another ensemble constant, $c_{I+1} = \langle \varphi_{I+1}(x) | \mathcal{E} \rangle$ and another Lagrange multiplier λ_{I+1}. The implications of theorems 5.2 and 5.3 is then that the class of probabilistic models derivable from maximizing entropy, subject to ensemble constants that are expectations of functions of the random variable x, is identical to the class of probabilistic models having sufficient statistics and consequently conjugate families of probability distributions for Bayesian updating of parameters. In view of the derivation of the entropy principle (theorem 5.1), these conjugate sampling models form a class identical to the class of ensembles characterized by a set of ensemble expectation constants plus the extended principle of insufficient reason. The implications of this important result will be examined in the next chapter.

5.5 Solutions for Probability Information

Before proceeding with these investigations, we might inquire if the exponential form of the probability distribution (5.35) is maintained if we allow other types of constraints on the ensemble. The exponential form is not maintained if we allow constraints that represent knowledge of fractiles of the distribution. Situations might arise in which it would be desirable to include testable information that is not in the form of an expectation or fractile of the distribution. (For example, (c) and (d) of (3.10)). However, we might expect that mixed expectation and fractile inequality constraints would constitute the most general case of practical interest in developing a probability distribution consistent with testable information. A solution for this case is easily developed using the standard Kuhn-Tucker conditions of non-linear optimization theory ([93]). The derivation will be given for a discrete space of N possible outcomes. However, the extension of the results to maximization of an entropy functional defined on a continuous outcome space (Jaynes [36],[40]) is possible using a more general formulation of mathematical programming ([32],[53]).

Consider the problem of maximizing

$$- \int_X p(x) \log p(x) \qquad (5.80)$$

subject to

$$p(x) \geq 0 \quad \text{all possible outcomes } x \qquad (5.81)$$

$$\int_X p(x) = 1 \qquad (5.82)$$

$$< \varphi_i(x) | \mathcal{S} > = \int_X \varphi(x) p(x) \leq c_i \qquad i = 1, \ldots, I \qquad (5.83)$$

$$\int_{x \leq x_j^*} p(x) \leq d_j \ , \quad j = 1, \ldots , J \ . \tag{5.84}$$

Constraints (5.81) and (5.82) state conditions necessary for $p(x)$ to be a probability distribution. The constraints (5.83) represent knowledge that the expected values of the functions $\varphi_1, \ldots , \varphi_I$ be less than or equal to given numbers $c_1, \ldots , c_I$. The constraints (5.84) represent knowledge that the cumulative distribution function evaluated at J points $x_1^*, \ldots , x_J^*$ is less than or equal to given numbers $d_1, \ldots , d_J$.

Since the constraints are linear in the probabilities $p(x)$, the Kuhn-Tucker conditions are necessary and sufficient for optimality. If λ_i, $i = 1, \ldots , I$, are the Lagrange multipliers corresponding to equations (5.83) and μ_j, $j = 1, \ldots , J$ the multipliers corresponding to (5.84), these conditions yield, for each outcome point x,

$$\log p(x) + \lambda_i \varphi_i(x) + \mu_j u(x_j^* - x) = a \tag{5.85}$$

where a is a constant arising from the normalization condition (5.82), and u is the step function

$$u(y) = \begin{cases} 1 & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases} .$$

In addition we have the complementary slackness conditions that

$$\lambda_i (\langle \varphi_i(x) | \mathcal{S} \rangle - c_i) = 0 \tag{5.86}$$

and

$$\mu_j \left(\int_{x \leq x_j^*} p(x) - d_j \right) = 0 \tag{5.87}$$

as well as the restriction that

$$\lambda_i \geq 0, \quad i = 1, \ldots , I, \quad \mu_j \geq 0, \quad j = 1, \ldots , J . \tag{5.88}$$

The form of the solution is clear from these expressions. Only those constraints that are binding (i.e., for which equality holds in (5.83), (5.84) contribute to the form of the solution for $p(x)$. Between any two adjacent fractile points x_j, $x_{j'}$, the distribution has the form of the exponential family. In passing through a fractile point x_j whose corresponding constraint (5.84) is binding the distribution $p(x)$ will change by a multiplicative constant.

As an illustration suppose that the outcome space is the set of integers $1, \ldots , N$, and the only information is that the median of the distribution is K. The maximum entropy distribution is then uniform on $1, \ldots , K$ and $K+1, \ldots , N$, with the values specified by

$$p(x) = \begin{cases} 1/2K & 1 \leq x \leq K \\ 1/2(N-K) & k < x \leq N \end{cases} . \tag{5.89}$$

As a second, somewhat more complex example, suppose that the outcome space is a set of evenly spaced positive real numbers. The mean is known to be m and the median is known to be g. Solution of the maximum entropy problem yields the following expression for the probability of an outcome point x:

$$p(x) = \begin{cases} \mu_1 e^{-\lambda x} & x \leq g \\ \mu_2 e^{-\lambda x} & x > g \end{cases} . \tag{5.90}$$

We shall make the usual assumption that the outcome points are sufficiently close together so that we can approximate sums over the outcome points by integrals and treat $p(x)$ as a probability density function. The unknown multipliers μ_1, μ_2, λ are then determined by the conditions:

$$\int_0^\infty p(x)dx = \int_0^g \mu_1 e^{-\lambda x}dx + \int_g^\infty \mu_2 e^{-\lambda x}dx = 1 \tag{5.91}$$

$$\int_0^g p(x)dx = \int_0^g \mu_1 e^{-\lambda x}dx = 0.5 \tag{5.92}$$

$$\int_0^\infty xp(x)dx = \int_0^g \mu_1 xe^{-\lambda x}dx + \int_g^\infty \mu_2 xe^{-\lambda x}dx = m \ . \tag{5.93}$$

Solving (5.91) and (5.92) for μ_1 and μ_2, we obtain

$$p(x) = \begin{cases} \dfrac{\lambda}{2(1-e^{-\lambda g})} e^{-\lambda x} & 0 < x \leq g \\ \\ \frac{1}{2} \lambda e^{-\lambda(x-g)} & g < x \end{cases} \ . \tag{5.94}$$

Then the equation (5.93) for the mean can be used to relate the multiplier λ to the mean m. Inserting (5.94) into (5.93), we find

$$\begin{aligned} m &= \frac{\lambda}{2}\left[\int_0^g \frac{x}{(1-e^{-\lambda g})} e^{-\lambda x}dx + \int_g^\infty xe^{-\lambda(x-g)}dx\right] \\ &= \frac{1}{2\lambda} \frac{(1 - 2e^{-\lambda g})(1 + g\lambda) + 1}{(1 + e^{-\lambda g})} \ . \end{aligned} \tag{5.95}$$

Numerical means would undoubtedly be required to find λ as a function of the known quantities m and g.

Chapter VI

INVARIANCE PRINCIPLES AS A BASIS FOR ENCODING INFORMATION

6.1 Decision Theory and Probabilistic Models

In the last section we developed theoretical foundations for structuring uncertainty based on two invariance criteria derived from the basic desideratum:

(1) exchangeability

(2) the principle of insufficient reason extended to sequences of outcomes.

The remainder of this dissertation will be devoted to exploring the implications of these invariance criteria to probabilistic modeling and statistical inference. The exploration will be brief and by no means comprehensive; much research remains to be done.

It is advisable to begin by reviewing the use of probabilistic models for encoding information in the context of decision theory. Expositions of decision theory usually deal with a conceptually simple form for the uncertainty: a probability distribution assigned to a set of mutually exclusive, collectively exhaustive outcomes. Each action available to the decision maker corresponds to a probability distribution assigned to this outcome space, and the optimum action is the alternative whose associated probability distribution yields the maximum expectation for the utility or value function assigned by the decision maker to these outcomes. Sequential decisions may be analyzed by using dynamic programming: The sequence of decisions and resolution of uncertainty may be structured as a decision tree, with the solution to be obtained by backward iteration.*

Decision tree methods have been applied to exceedingly complex problems in sequential decision-making (for example, [56]), but the applicability of the method is severely limited by the requirement that all decision alternatives and possible outcomes be specified in advance. When repeated decisions or complex information-gathering alternatives are included, the decision trees can easily exceed the proceeding capabilities of the largest computers.

The alternative to the decision tree approach is the use of probabilistic models. A considerable literature exists on probabilistic models from the viewpoint of Bayesian decision theory ([2],[49],[70]). Much of the emphasis in this literature is directed at the differences between Bayesian and classical approaches, and little attention has been paid to the fundamental justification for the models.

Our basic assumption that probability theory is a means of representing information leads immediately to the idea that information is encoded in the choice of a probabilistic model. The invariance approach allows us to understand the correspondence between states of information and many of the commonly used models of probability theory.

The transition to probabilistic models is effected by dropping the requirement that the uncertainty concerns a single set of mutually exclusive and collectively exhaustive outcomes. We may assume that

* It is assumed that the reader is familiar with these ideas, developed in detail in such references as [48], [68], [70].

the uncertainty possesses a repetitive structure. Suppose we have a sequence $[x_t]$ of uncertain quantities defined on the same space of possible outcomes. For convenience we will refer to these sequence elements as "observations," although this term may not be appropriate in all cases. Usually the parameter t indexing sequence elements will correspond to time.

Allowing different probability distributions to be assigned to different elements in the sequence $[x_t]$ provides a general framework for modeling uncertain repetitive processes. In some decision situations we may wish to include an additional uncertainty concerning the outcome that follows a terminal decision, because the value structure that constitutes the decision criteria depends on both this terminal outcome and the sequence $[x_t]$. The framework is then equivalent to the formulation of decision theory proposed by Raiffa and Schlaifer ([70], Chapter 1); our formulation stresses the repetitive nature of the experimentation that might be associated with the sequence $[x_t]$.

We shall not take the time to develop applications of the formulation. This straightforward task is left to the reader, whom we presume to be familiar with the literature on probabilistic models in management science. Even a cursory review of this literature would require a substantial deviation from our purpose here, which is to provide a unified viewpoint on probability theory as a means of encoding information. We might point out that the repetitive structure $[x_t]$ might be applied whenever decisions are made on the basis of repetitive observations. Examples of such decision problems include:

(1) estimation of the parameters of a population or selection of a terminal act (i.e., "accept" or "reject"), on the basis of a sequence of sample observations. Such problems have been extensively examined in the literature, and the Bayesian viewpoint is summarized in such books as [49], [70].

(2) sequential decision and control problems, such as equipment maintenance, inventory replenishment, production process control, and choice of competing processes, in which an ongoing series of decisions must be considered. An introduction to the literature on these problems may be found in such references as [2], [24], [67], [76].

6.2 Stationary Processes: Exchangeable Sequences

In the most general case we might allow the probability distribution assigned to each observation in the sequence $[x_t]$ to be different. Such a situation could occur only if the decision-maker has considerable knowledge about the process that is generating the x_t. It places a considerable burden on the decision-maker to encode this knowledge in the probabilistic form needed for a quantitative analysis.

We shall begin by examining the situation in which the decision-maker is comparatively ignorant. In particular, we shall consider the decision-maker's state of knowledge to be so limited that he perceives no distinction between the observations. There is no relevant difference between his information about any group of n

elements in the sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ and any other group of n elements, $x_{t'_1}, x_{t'_2}, \dots, x_{t'_n}$, where $n = 1, 2, 3, \dots$. The sequence $[x_t]$ then constitutes a sequence of exchangeable trials.

The characterization of the sequence $[x_t]$ as exchangeable implies two properties:

(1) stationarity: the probability distribution assigned to any group of n observations $x_{t_1}, \dots, x_{t_n}$ is invariant to a change in all of the parameter values by some fixed amount h.

(2) conditional independence of the observations, as implied by the de Finetti theorem, (5.1). The parameters $t_1, \dots, t_n$ serve only as identifying labels; the probability assignment to any group of n observations must be invariant under permutation of these the labels. Henceforth, we will simplify the notation by using integer labels for the observations: $x_1, x_2, \dots, x_n$.

We shall examine ways of relaxing these two assumptions later in this chapter.

Conditional independence implies strong limitations on the inferences that can be drawn from a sequence of observations $x_1, \dots, x_n$. Knowledge of some observations affects the probability distribution assigned to other observations only by revising the state of knowledge regarding the histogram ω summarizing an arbitrarily large number of observations, as we saw in the early part of Chapter 5.

If the $[x_t]$ process may be characterized as exchangeable, the uncertainty about the process is equivalent to the uncertainty about the histogram ω. If we know the histogram ω, then any group of observations $x_1, \ldots, x_n$ may be regarded as independent, identically distributed random variables whose probability distribution corresponds to the histogram ω. The simplification over a decision tree approach is that we do not need to consider all the different possible ways of achieving the same histogram. From the exchangeability assumption the different ways are judged to be equally probable. *The essence of inference on an exchangeable sequence is the characterization of the histogram ω.* It is this distribution that provides the likelihood function $\{x|\omega,\mathcal{E}\}$ that is the fundamental element of Bayesian (and classical) statistics. The histogram ω is the uncertain outcome for a process generating exchangeable observations $[x_t]$. Knowledge of ω might be considered statistical perfect information: the situation in which the probability distribution is "known" in the usual sense of classical statistics. Further information that changes the probability distribution assigned to an observation or a group of observations would violate the exchangeability assumption, but of course inference is still possible using Bayes' Rule if a conditional probability structure has been assigned.

The difficulty inherent in developing a general formalism for inference on an exchangeable process is the dimensionality of the outcome space. Unless the set of possible values that the observations x_j may assume is limited to a small number of discrete points, the dimensionality of the possible histogram ω becomes unmanageable.

We need a means of reducing the dimensionality of ω to a level consistent with available analytic resources. There are two possibilities. The first is the postulation of a "model space," i.e., a set of possible histograms ω_d for $d \in D$ of which one is assumed to generate the data. Such an approach has been discussed by Murray and Smallwood [58], [59], [79] and Matheson [55]. The observation $x_1, \ldots , x_n$ are used to update a probability distribution on which histogram represents the "true" distribution that will be achieved by an arbitrarily large number of observations.

The difficulty with such an approach lies in defining a suitable model space. The choice of a set of possible histograms determines the level of complexity, hence the cost of the analysis, and this choice implicitly encodes subjective information by eliminating other possible histograms on the outcome space. It would seem desirable to have general principles for the selection of a "model space" rather than being forced to rely on an ad hoc procedure.

Another means of reducing the dimensionality of the histogram ω to a manageable level derives from the basic desideratum. The extended principle of insufficient reason permits us to overcome the dimensionality problem in assigning a likelihood function if we can represent the relevant information by a small number of constraint relations on the histogram ω.

In the limit of a large number of observations, one histogram will result from more outcome sequences than other possible histograms by an arbitrarily large margin. The problem of finding this special histogram may be solved by maximizing the entropy functional of the probability distribution subject to the given constraints, as we saw in Chapter 5. From the assumptions of exchangeability and the extended principle of insufficient reason, this histogram will be realized with probability one in a large enough sequence $[x_t]$.

The extended principle of insufficient reason for exchangeable sequences provides a direct link between states of information and many familiar probability distributions. This link may act in either direction. For a given state of information the principle will specify which probability distribution is appropriate for $\{x | \omega, \mathcal{E}\}$. The link may be used in the other direction when a specific form is proposed as the likelihood function in a problem involving repeated observations. By using the extended principle of insufficient reason we can determine the state of information that corresponds to this probabilistic model, and we can then ascertain if this state of information really reflects what the decision-maker believes.

Let us now examine some specific models, using the results of Chapter 5. Solutions of the maximum entropy problem for various probabilistic constraints of the expectation type were tabulated in table 5.1, and these have a clear interpretation in terms of the extended principle of insufficient reason. We have already examined one example in Chapter 5. If the outcome space of the observations x_t lies in the set of non-negative numbers, and the state of information is that all sequences having the same mean are equally probable, then the extended principle of insufficient reason indicates that the exponential distribution is the appropriate likelihood function for this exchangeable sequence.[*]

Let us examine the link from the other direction: if an exponential distribution is used for the likelihood function, then this model implies that all sequences having the same mean are equally likely.

This viewpoint on the exponential distribution will be familiar to many readers; it is in fact a basis for the axiomatic derivation of the Poisson process[**] that is equivalent in this special case to the general maximum entropy derivation in Section 5. For other distributions the reasoning is equivalent but not as familiar. If sequences with the same mean and variance are judged equally probable, the appropriate likelihood function is the normal. This result gives us an insight into the meaning of the central limit theorem; the normal distribution corresponds to a limiting state of information in which only knowledge of the mean and variance of a random variable remain. If sequences having the same mean and geometric mean are judged equally probable, the corresponding likelihood function is a gamma distribution.

This list can be continued to generate many commonly used distributions. Other distributions such as the Weibull and Cauchy may be generated by a transformation of variables from a maximum entropy distribution such as the exponential or uniform (Tribus, [85]). As

[*] We avoid the difficulties of inherent in defining entropy on a continuous probability space by finding the maximum entropy distribution for a discrete space of N outcome points, then letting N go to infinity and approximating the sum over possible outcomes in the partition function as an integral. See the footnote of page 81.

[**] A Poisson process is a counting process with exponentially distributed interarrival times. The axiomatic derivation and the characterization in terms of the distribution on interarrival times may be found, for example, in Parzen [62].

we remarked in Chapter 5, the Bernoulli case is special because the outcome space has only two points. The histogram specifying the results of a large number of trials has only one degree of freedom, so the extended principle of insufficient reason is not necessary. The assumption that the observations $\{x_t\}$ form an exchangeable sequence specifies the model up to the value of a single parameter.

The theorems (5.2) and (5.3) imply that if the extended principle of insufficient reason is applied subject to expectation constraints,

$$< \varphi_i(x_t)|\mathcal{E}> = c_i \qquad i = 1, \ldots , I \tag{6.1}$$

the resulting probability distribution has additive sufficient statistics. Moreover, if a probability distribution has sufficient statistics and known boundaries delineating the regions of zero probability, then the sufficient statistics may be written in the additive form

$$s_i(x_1, \ldots , x_n) = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j) \qquad i = 1, \ldots , I . \tag{6.2}$$

The number of observations n is a known auxiliary statistic, and the equivalence classes for the statistics are clearly unchanged if we use the form (6.2) instead of (5.59). From the results of theorems (5.2) and (5.3) we can characterize the state of information that leads to this probability distribution using the extended principle of insufficient reason: All sequences of observations that lead to the same value for the sufficient statistics s_i, $i = 1, \ldots , I$, in (6.2) must be equally probable. But s_i is the sample average of the function φ_i. By the strong law of large numbers,

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j) &= s_i(n) \\ &= c_i = < \varphi_i(x_t) | \omega , \mathcal{E} > \end{aligned} \tag{6.3}$$

as $n \to \infty$. We may regard this result as the basic ergodic theorem for exchangeable sequences. The equivalence between the limit of sufficient statistics that are sample averages and the expectation over the probability distribution corresponding to ω gives us a fundamental insight into why expectation information is important. In the case of fractile constraints, it does not appear to be possible to summarize the content of a group of observations by sufficient statistics except

(1) in the case where the fractiles are 0 or 100 per cent; i.e., they define the boundaries for regions of zero probability, or

(2) when the number of observations is essentially infinite.

The extended principle of insufficient reason reduces the problem of inference on an exchangeable sequence from inference on the histogram ω to inference on the value of parameters c_i of equation (6.1). Viewing the situation from the other direction, it is the fact that all sequences having the same (sample) average value for the functions $\varphi_i(x_t)$ are equally probable that allows the inference on the parameters c_i to proceed in such a simple fashion: we need only keep a record of the sample averages (6.2): these averages are sufficient statistics for inference on the parameters c_i. Any inference involving sufficient statistics is equivalent to this process, excepting the case in which the inference is on the boundaries of a region of zero probability in the outcome space.

An exchangeable sequence characterized by the extended principle of insufficient reason subject to a set of expectation constraints (6.1) is then specified by the current state of knowledge concerning the parameters c_i, $i = 1, 2, \ldots, I$. These parameters are defined operationally through equation (6.3); that is, they are interpreted as an average over the frequency distribution that would result from a long sequence of observations $[x_t]$. The probability distribution $\{c_1, \ldots, c_I|\mathcal{E}\}$ on the possible values of these parameters, assigned on the basis of a state of information $\mathcal{E}$, completes the specification of probabilistic structure.

To recapitulate, knowledge of the $[x_t]$ process has been encoded in the following:

(a) assumption that any finite sequence drawn from the $[x_t]$ process is an exchangeable sequence

(b) assumption of the extended principle of insufficient reason: All sequences having the same sample average values of the functions $\varphi_i(x_t)$, $i = 1, \ldots, I$, are equally probable

(c) the probability distribution $\{c_1, \ldots, c_I|\mathcal{E}\}$ assigned to the limiting values

$$c_i = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_{t_j}) \ . \qquad (6.5)$$

The literature in the Bayesian school of statistics has concentrated almost exclusively on stage (c) of the encoding process, the encoding of prior knowledge on distribution parameters, while (a) and (b) have been implicitly assumed. The distribution

$$f(c_1, \dots , c_I) = \{c_1, \dots , c_I|\mathcal{E}\}$$

that reflects the initial prior state of information may be quite arbitrary, but if subsequent information is strictly in the form of observations from the process, then the family of possible posterior distributions reflecting the state of information after the observations have been assimilated may be written in a simple form. Since from theorems 5.2 and 5.3 the likelihood function for n observations must be of the form

$$\begin{aligned}&\{x_1, \dots , x_n|c_1, \dots , c_I,\mathcal{E}\}\\ &= Z^{-n} \exp\left[\sum_{i=1}^{I} n\lambda_i s_i(x_1, \dots , x_n)\right]\end{aligned} \tag{6.6}$$

where $s_i(x_1, \dots , x_n) = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j)$ is the sample average of the function φ_i, $i = 1, \dots , I$. The distribution depends on $c_1, \dots , c_I$ through the functions $\lambda_i = \lambda_i(c_1, \dots , c_I)$ and the normalization function

$$Z = Z(\lambda_1, \dots , \lambda_I) = \int_x \exp\left(\sum_{i=1}^{I} \lambda_i\varphi_i(x)\right) . \tag{6.7}$$

Since $c_i = < \varphi_i(x)|\mathcal{E} >$, we see that

$$c_i = \frac{\partial}{\partial\lambda_i} \log Z(\lambda_1, \dots , \lambda_I) \tag{6.8}$$

giving the $c_1, \dots, c_I$ in terms of the functions $\lambda_1, \dots, \lambda_I$, or alternatively, these relations (6.8) can be used to determine the $\lambda_1, \dots, \lambda_I$ in terms of the $c_1, \dots, c_I$.

In applying Bayes' rule to revise the probability distribution on the constants $c_1, \dots, c_I$ the normalization function cancels. If the likelihood function is of the form (6.6), then the posterior distribution will be of the form

$$\{c_1, \dots, c_I | x_1, \dots, x_n, \mathcal{E}\} \\ = (Z^*)^{-1} f(c_1, \dots, c_I) \exp\left(\sum_{i=1}^{I} n\lambda_i s_i(x_1, \dots, x_n) \right) \tag{6.9}$$

where Z^* is a new normalizing function determined by the requirement that the total probability for all values of the outcome space of the $c_1, \dots, c_I$ be equal to one.

Since the functions $\lambda_i(c_1, \dots, c_I)$ do not depend on the observations, the additive statistics

$$s_i(x_1, \dots, x_n) = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j)$$

and n, which are of fixed dimensionality, specify the posterior distribution. It is sometimes convenient to think of n and the $s_i(x_1, \dots, x_n)$ as defining a parameter space, each point of which corresponds to a posterior distribution. The effect of more observations $x_{n+1}, x_{n+2}, \dots$ is to cause transitions to new points in the parameter space; these transitions are stochastic in the sense that the posterior values of the statistics s_i are not known prior to the observations. This approach of using a parameter space defined by the sufficient statistics (relative to a given prior distribution $\{c_1, \ldots ,c_I|\mathcal{E}\}$ has been widely applied to problems involving sequential sampling decisions (Bather [3], Bather and Chernoff [4], Chernoff [7], Chernoff and Ray [8], Lindley [48], Lindley and Barnett [50], Pratt [64], Wetherill [91]).

6.3 Prior Probability Assignments to Parameters

A remaining question is that of encoding the original prior probability distribution $\{c_1, \ldots ,c_I|\mathcal{E}\}$ on the parameters $\{c_1, \ldots ,c_I|\mathcal{E}\}$ corresponding to the long-run averages of the functions $\varphi_1, \ldots , \varphi_I$:

$$c_i = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j) \ . \qquad (6.10)$$

Much has been written on this question of encoding probability distributions on distribution parameters, and we do not propose to summarize this voluminous literature here. The methods for encoding a probability distribution on a set of uncertain quantities apply here in the same way as any other; the conceptual basis for the encoding is the operational definition (6.10) of the parameters $\{c_1,\ldots,c_I|\mathcal{E}\}$ in terms of a long-run frequency average computed from exchangeable observations.

There are certain special situations in which it is possible to apply invariance principles directly to the encoding of the prior distribution $\{c_1, \ldots ,c_I|\mathcal{E}\}$ on the distribution parameters $c_1, \ldots ,c_I$. This use of invariance principles is due to Jaynes [40], and in large part it has motivated the present work. Suppose that the decision-maker's state of information has the following property. If the problem is transformed into a new set of variables x^*, $c_1^*, \ldots, c_I^*$ according to some given set of transformation equations, the decision-maker is unable to distinguish the transformed problem from the original problem. His state of information is the same for both the original and the transformed problem, and therefore by the basic desideratum he should assign the same probability distribution in both cases.

An example will help to clarify the reasoning involved. Suppose that we have two constants

$$c_1 = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} x_j = < x|\mathcal{E} > \tag{6.11}$$

$$c_2 = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} x_j^2 = < x^2|\mathcal{E}> . \tag{6.12}$$

The distribution on the exchangeable observations $x_1, x_2, \ldots$ is specified by these constants and the extended principle of insufficient reason; any two sequences of observations having the same mean and mean square are judged equally probable. For convenience it is useful to use a parametrization in terms of a location parameter μ and a scale parameter σ:

$$\begin{aligned} \mu &= c_1 \\ \sigma &= \sqrt{c_2 - c_1^2} \quad \text{or} \quad \sigma^2 + \mu^2 = c_2 \end{aligned} \tag{6.13}$$

Clearly a joint probability distribution on c_1 and c_2 implies a joint probability distribution on μ and σ and visa versa, since one may transform from one set of independent variables to the other using well-known techniques for change of variables.

Then we will take our problem to be that of specifying a joint probability distribution $\{\mu,\sigma|\mathcal{E}\}$ on the location parameter μ and the scale factor σ, where $\sigma^2 = \langle (x - \langle x|\mathcal{E}\rangle)^2|\mathcal{E}\rangle$.

Suppose that we are ignorant of μ and σ in the sense that any linear transformation on x yielding new parameters μ^* and σ^* leads to exactly the same state of information: the decision maker cannot distinguish between the two problems. If

$$\begin{aligned} \mu^* &= \mu + b \\ \sigma^* &= a\sigma \\ x^* - \mu^* &= a(x - \mu) \end{aligned} \tag{6.14}$$

then in order for the prior distributions to be the same for both the original and the transformed problem, the probability density function $f(\mu,\sigma)$ must satisfy the functional equation

$$f(\mu,\sigma) = af(\mu + b, a\sigma) \ . \tag{6.15}$$

The solution to this equation is ([24])

$$f(\mu,\sigma) = \frac{\text{constant}}{\sigma} \tag{6.16}$$

which we may take to correspond to a state of "complete ignorance" about the zero point μ and scale factor σ; this distribution remains invariant to an arbitrary change in zero point and scale factor. This form of a prior distribution to represent complete ignorance of a zero point and scale factor was initially proposed by Jeffreys [43].

6.4 More Complex Applications of Exchangeability

In the preceding sections we have developed a framework for encoding information about an uncertain sequence of outcomes, $[x_t]$, providing that the individual sequence elements were exchangeable. We noted that exchangeability is an invariance concept that proceeds from the basic desideratum. If permuting the sequence elements does not change our state of information then it should not change the joint probability distribution that we assign to these sequence elements. The limit of perfect information about an exchangeable process corresponds to knowing the histogram of observed outcomes for a large number of sequence elements. The probability distribution assigned to any unknown sequence element x_t is equal to this histogram in the limiting case of perfect information. In this limit the probabilistic models derived from exchangeable sequences are equivalent to sequences of independent, identically distributed random variables. Without perfect information we must deal with the problem of inferring this histogram that is equivalent to the probability distribution for the independent random variables.

Exchangeability can provide the conceptual basis for probabilistic models in other situations besides the case of independent, identically distributed random variables. Continuous random processes, time varying processes, and Markov-dependent processes may all be analyzed in terms of exchangeability. We shall now sketch very briefly how the exchangeability concept could be generalized to these more complex situations.

Let us first consider the case of a continuous observation process $[x_t]$. A simple generalization of the exchangeability concept is to apply it to increments of the $[x_t]$ process rather than the process itself. A derivative process $[y_t(h)]$ is judged to be exchangeable for any fixed value h, where the sequence elements of the derivative process are computed from the original process $[x_t]$ by the relation

$$y_t(h) = x_{t+h} - x_t \ . \tag{6.17}$$

If the index set t for the $[x_t]$ process is allowed only to take discrete values, constructing an incremental process $y_t(h)$ is a trivial extension of our previous analysis. For a continuous observation process $[x_t]$ exchangeable increments is a strong assumption with strong consequences. Exchangeability must hold for increments of any size: for $h > 0$, any sequence of non-overlapping increments $[y_t(h)]$ must be exchangeable. Likewise, a sequence of non-overlapping increments $[y_t(h/n)]$ must also be exchangeable, for any integer $n = 1, 2, 3, \dots$.

Let us now examine the consequences of de Finetti's theorem. Both of these sequences must be composed of random variables that are conditionally independent: A histogram constructed from sequence elements becomes equivalent to the probability distribution assigned to one of these exchangeable elements. Then an increment $y_t(h)$ can be written as

$$y_t(h) = \sum_{j=o}^{n-1} y_{t+jh/n}(h/n) \tag{6.18}$$

where the $y_{t+jh/n}$ are independent and identically distributed random quantities. The values of n and h may be chosen arbitrarily, and we may draw the following conclusion. For any integer n, an increment of the process may always be written as the sum of n independent and identically distributed random quantities. A random variable possessing the property above is said to be <u>infinitely divisible</u>.

Two examples will illustrate continuous observation processes $[x_t]$ whose increments are infinitely divisible, the Wiener (or normal) process and the Poisson process. For the Wiener process $y_t(h)$ will be normally distributed with mean zero and variance $\sigma^2 h$, proportional to the increment size h; $y_t(h)$ is equal to the sum of n independent, identically distributed normal random variables with mean zero and a variance of σ^2/nh. We shall consider a general Poisson process in which x_t undergoes discrete jumps of a fixed but arbitrary size u. (The usual case of the Poisson process is $u = 1$.) If ν denotes the intensity, the mean number of jumps per unit time, then the probability distribution assigned to an increment of the process is

$$\{y_t(h) = ju|\mathcal{E}\} = \frac{(\nu h)^j}{j!} e^{-\nu h} . \qquad (6.19)$$

The increment $y_t(h)$ is equivalent to the sum of n independent and identically distributed (Poisson) random variables, each having discrete jumps of size u occurring with an intensity of ν/n per unit time.

The most general case of infinite divisibility consists of a sum of (independent) Wiener and Poisson processes. If a continuous process has infinitely divisible increments, it may be written as a sum

$$x_t = \gamma t + n_t + \sum_u z(u) \qquad (6.20)$$

a fixed trend of γ per unit time, a Wiener process n_t, and a sum of Poisson processes having various (positive and negative) jump sizes u. This characterization of infinitely divisible processes is very strong. We need only specify the fixed trend γ per unit time, the drift rate σ^2 of the Wiener process, and the intensity $\nu(u)$ as a function of the jump size of the Poisson process components, and we have specified the $[x_t]$ process completely.

There is a considerable literature in advanced probability theory and stochastic process theory on infinite divisibility, and it is possible to make the foregoing assertion on the general form of an infinitely divisible process precise and rigorous. The assertion is proved using a representation theorem for the characteristic function of an arbitrary increment $y_t(h)$. The rather arduous details may be found in Loève [51], sections 22 and 37, Doob [12], or Gnedenko and Kolmogorov [25]. Our interest enters on the fact that this area of the literature of probability theory may be related to an invariance principle, for as we have seen, infinite divisibility is immediately implied by the concept of exchangeability for the increments of a continuous process.

More general applications of exchangeability rely on the following extension of the exchangeability concept, suggested by de Finetti [21]. Consider a sequence of uncertain quantities $[x_t]$ defined over the same outcome space. We shall define the sequence to be conditionally exchangeable if these quantities can be sorted into disjoint classes $\alpha_1, \alpha_2, \ldots$ in such a way that the sequence elements $x_{t_1}, x_{t_2}, \ldots$ within each class are exchangeable. For an exchangeable sequence all sequence elements must be exchangeable, but for a conditionally exchangeable sequence it is only necessary that the sequence elements in each class be exchangeable, i.e., the joint probability distribution assigned to these elements must be invariant to any arbitrary permutation of the elements within a class.

One of the simplest applications of conditional exchangeability is to time-dependent models. The elements x_t of an observation process $[x_t]$ are sorted into classes according to the time index t. A coin tossing process in which two different coins are tossed alternately illustrates how conditional exchangeability can be applied to cyclical or seasonal processes. Tosses of the first coin would occur on odd-numbered trials, and tosses of the second on the even trials; the classes whose elements are judged to be exchangeable are then the odd numbered trials and the even numbered trials. The inference problem concerns the histograms for odd and for even tosses. Each histogram is determined by one number, the long-run fraction of heads, and knowledge of this fraction for odd and even tosses would constitute perfect information for the process.

Instead of being determined a priori the assignment of sequence elements to the exchangeable classes may depend on the outcomes of selected sequence elements. Renewal processes are of this type. An example might be machine replacement, in which a machine is observed to work or to fail at each period t. If machines that fail are replaced by new machines, the time since the last renewal (e.g., the age of the machine) constitutes the criterion for sorting machine performance data into exchangeable classes.

If the exchangeable classes are defined by the observation immediately preceding, then conditional exchangeability forms a natural basis for Markov models. For a discrete time index observations $x_{t_1}, x_{t_2}, x_{t_3}, \ldots$ will be judged exchangeable and only if the preceding outcomes $x_{t_1-1}, x_{t_2-1}, x_{t_3-1}, \ldots$ are all equal. The inference problem for such a Markov model involves a separate histogram for each possible outcome of a sequence element x_t; the set of such histograms determines the transition probability matrix that is the focus of the analysis of Markov processes.

More complex Markov models can be constructed by taking more of the history of the process into account in setting up the exchangeable classes. Breaking an exchangeable class into two separate classes is equivalent to splitting a Markov process state into substates. Likewise, merging the elements of several exchangeable classes into a single class is equivalent to coalescing several Markov process states together into one state.

The concept of exchangeability leads naturally to a partial ordering among models. At one extreme is the situations in which all sequence elements x_t are considered exchangeable; there is only one class and the decision-maker's state of information is invariant to any permutations of elements. If the decision-maker compares the original sequence $[x_t]$ with a rearrangement and finds that they do not correspond to equivalent states of information, then at least two classes of exchangeable elements are required. At the other extreme is where no permutation of sequence elements will result in an equivalent state of information.

A limiting case of a conditionally exchangeable sequence occurs when each observation x_t in a sequence $[x_t]$ must be placed in a separate class, so no two observations are exchangeable. Exchangeability is then a vacuous concept, since observations cannot be summarized by histograms; there are no long-run fractions. The encoding process must consist of assigning probabilities to each combination of possible outcomes for the sequence $[x_t]$. One is forced to use a tree structure or its equivalent, with probabilities assigned separately to each outcome in the sequence, conditional on all of the outcomes that have preceded it. The exchangeability concept will only have value when many information-equivalent permutations of sequence elements exist, and it is possible at least conceptually to consider the possible histograms that might result from a large number of sequence elements in one class.

Whereas for exchangeable sequences the inference problem deals with the histogram that would result in the limit of a large number of sequence elements, for a conditionally exchangeable sequence a separate histogram will be constructed for each class $\alpha_1, \alpha_2, \ldots$ and the inference problem will concern these histograms jointly: of all the possible combinations of histograms, ω_1 for α_1, ω_2 for $\alpha_2, \ldots$, etc., which set of histograms will result from a long series of observations of the process. This joint inference problem is likely to be extremely formidable unless it is possible to separate the information that pertains to each class α_i, that is, if observations in class α_i only change our state of knowledge about the long run histogram that will result from a large number of observations in class α_i. Further, information about the histogram that results from a large number of observations of class α_j, $j \neq i$, will not be changed by observing one or a number of sequence elements in class α_i. When such a separation of information is possible inference on a conditionally exchangeable sequence becomes equivalent to a great many inference problems on exchangeable sequences, one for each class α_i. The extended principle of insufficient reason may be used to reduce the dimensionality of the inference problem to a manageable level, as we have discussed in Section 6.2 and in Chapter 5.

Unfortunately, the inference problems that are actually encountered may not possess this separation property. By flipping one coin we may change our assessment of the limiting fraction of heads for a second coin. Methods for dealing with such joint inference problems should be a goal for future research.

Chapter VII

TESTING INVARIANCE - DERIVED MODELS

7.1 The Form of the Test

In previous chapters we have seen how invariance principles may be used to generate specific probabilistic models. The basic desideratum that equivalent states of information must lead to the same model forms the foundation for the invariance approach. A model summarizes relevant information in explicit form, and as the state of information changes over time it may be necessary to revise the models that summarize this information.

The invariance approach to probabilistic encoding of information is based on the decision-maker's assessment of equivalence between states of information. The Ellsberg urn examples of Chapter 2 provide simple situations in which this equivalence seems intuitive. Many other examples exist in which the equivalence is based on physical symmetry, e.g., a fair coin, fair die or wheel of fortune. Our strength of belief in this equivalence may vary from one situation to the next, and in particular we may question an equivalence that was originally assumed after additional information has been received. In order to make proper use of the invariance approach to probabilistic encoding we should have a means of revising the original invariance assumptions as more data becomes available.

Additional information rarely provides grounds for stating that a probabilistic model is wrong. Such an inference only follows if the information was judged impossible a priori on the basis of the model. Usually information is judged more or less improbable, and even a very improbable result cannot be said to invalidate a model.

Perhaps the best way to approach the difficult problem of testing invariance - derived models it to examine a simple case. Suppose that we are given a coin presumed on the basis of our experience to be "fair." We toss this coin twenty times and observe the result, which might be

(a) T T T T H T T H H T T H T H H H H T T H

(b) H H H H H H H H H H H H H H H H H H H H

Result (a) will not lead us to question our model that the coin is fair, while with result (b) we will feel that we have strong grounds for questioning whether tosses of this coin are really Bernoulli trials whose limiting long-run fraction of heads is one-half. Yet on the basis of the assumed model, both results (a) and (b) would have been assigned the same a priori probability, namely, one chance in 2^{20}, or about 10^{-6}. It is the fact that the outcome (b) of twenty straight heads is so much more probable than outcome (a) on the basis of alternate models (e.g., someone has given us a two-headed coin) that leads us to question whether the fair coin model is adequate to represent the new state of information.

The fair coin represents about the simplest conceivable invariance model. The assumption that tosses of the coin form a sequence of exchangeable trials follows from the fact that no causal relationship is perceived that would allow the results of one toss to influence another, and the properties of the coin-tossing mechanism are assumed to stay constant over time. The symmetry of the coin motivates the application of the criterion of insufficient reason in assigning the probability of heads (equal to the expected value for the long-run fraction of heads). We have no reason to distinguish between the two possible outcomes of heads and tails, therefore our state of information is unchanged if these outcome labels are interchanged, and we must assign to heads a probability of one half. For other invariance models such as those mentioned in Chapter 6 we can proceed through a similar chain of reasoning. We shall develop the concepts we need to test the invariance assumptions of the "fair coin" model. The extension to the case of more general invariance models is straightforward.

To reexamine the "fair coin" model based on the additional information gained by observing the results of twenty flips we proceed in a familiar way, starting with Bayes' Rule. Suppose we had assigned a prior probability of $\{m|\mathcal{E}\}$ to the event that the fair coin model is correct (that is, any subsequence of tosses selected prior to knowing the results would have a limiting fraction of heads that approaches one half). The likelihood function for any sequence of n tosses of a fair coin is

$$\{E|m,\mathcal{E}\} = (\tfrac{1}{2})^n \; . \tag{7.1}$$

In order to proceed further we need to specify the possible alternatives that could occur if the fair coin hypothesis proves false. It is this complication that causes most of the difficulty. Suppose we could specify some alternative model $m*$ and require that either the fair coin model m or the alternative $m*$ hold true:

$$\{m|\mathcal{E}\} + \{m^*|\mathcal{E}\} = 1 \; . \tag{7.2}$$

The likelihood function $\{E|m^*,\mathcal{E}\}$ is presumed to be a known function of the data. For example, if m^* were the model that the coin has two heads, we would have

$$\begin{aligned} &\{E = n \text{ heads in } n \text{ tosses}|m^*,\mathcal{E}\} = 1 \\ &\{E = \text{any other result} \qquad |m^*,\mathcal{E}\} = 0 \; . \end{aligned} \tag{7.3}$$

Now it is a simple matter to use Bayes' Rule to revise our prior probability assignments on the two models m and m^* to reflect new data E:

$$\{m|E,\mathcal{E}\} = \frac{\{E|m,\mathcal{E}\}\{m|\mathcal{E}\}}{\{E|m,\mathcal{E}\}\{m|\mathcal{E}\} + \{E|m^*,\mathcal{E}\}\{m^*|\mathcal{E}\}} \tag{7.4}$$

$$\{m^*|E,\mathcal{E}\} = \frac{\{E|m^*,\mathcal{E}\}\{m^*|\mathcal{E}\}}{\{E|m,\mathcal{E}\}\{m|\mathcal{E}\} + \{E|m^*,\mathcal{E}\}\{m^*|\mathcal{E}\}} \tag{7.5}$$

$$= 1 - \{m|E,\mathcal{E}\} \; . \tag{7.6}$$

If the original state of information led to the probability assignment of one chance in a million that the coin was not fair, the posterior probability that the coin is not fair is about 50 per cent.

The problem with this analysis is that it ignores the multitude of ways in which the fair coin hypothesis might be violated. The limiting fraction of heads might be somewhere between 0.5 and 1.0, or perhaps the flips cannot be considered as independent trials and a Markov model is needed to describe their probabilistic behavior.

We would like therefore to eliminate the normalization requirement (7.2) from our analysis. We can accomplish this objective by choosing to work with ratios of probabilities. Dividing (7.4) by (7.5) yields,

$$\frac{\{m|E,\mathcal{E}\}}{\{m^*|E,\mathcal{E}\}} = \frac{\{E|m,\mathcal{E}\}\{m|\mathcal{E}\}}{\{E|m^*,\mathcal{E}\}\{m^*|\mathcal{E}\}} \; . \tag{7.7}$$

This quotient form of Bayes' Rule avoids the need for a normalization. We can speak of the odds in favor of one hypothesis compared to another without reference to additional hypotheses that might also be possible. In our example, we could say that a fair coin was a million times more probable than a two-headed coin, but after observing twenty straight heads the posterior probability assignments are approximately even.

We may gain some additional insight into the implications of the observed data on the comparison between two alternative models if we change (7.7) from a multiplicative to an additive form by taking logarithms. Good [27], Jaynes [35] and Tribus [83], [85] have suggested using base 10 logarithms, multiplied by ten to achieve a scale that is intuitive for many engineers. The logarithm of the odds is defined to be the <u>evidence function</u>

$$\mathrm{ev}(m{:}m^*|\mathcal{E}) = 10 \log_{10} \frac{\{m|\mathcal{E}\}}{\{m^*|\mathcal{E}\}} \tag{7.8}$$

and is measured in decibels (db). A table for conversion between odds and evidence in decibels is found in [85]. Using the evidence function, the quotient form of Bayes' Rule (7.7) may be written as

$$\mathrm{ev}(m{:}m^*|E,\mathcal{E}) - \mathrm{ev}(m{:}m^*|\mathcal{E}) = 10 \log \frac{\{E|m,\mathcal{E}\}}{\{E|m^*,\mathcal{E}\}} \; . \tag{7.9}$$

The difference between the posterior evidence and the prior evidence for one hypothesis as opposed to another is the logarithm of the likelihood ratio (multiplied by ten in order to have the units be decibels).

Before we can use (7.9) we must select the model m and an alternative m^*. Since the model space M may be large, the selection of the alternative model m^* poses a difficulty. We adopt a worst case approach. This approach to the problem of selecting an alternative model has been suggested by Jaynes [41] and Tribus [83], [85]. It is useful when the model m is derived from invariance assumptions that imply equivalence between probability assignments; relaxing these equivalences may permit a large set of possible alternative models.

Suppose we choose m^* so that the difference between the prior odds and the posterior odds for m compared to m^* is made as large as possible. Equivalently, we could choose m^* to give the prior evidence function minus the posterior evidence function its maximum value. This choice is equivalent to making the log likelihood ratio in (7.9) a minimum, or, since m^* only enters the denominator, to make the likelihood function $\{E|m^*,\mathcal{E}\}$ a maximum by choosing the proper model m^* in the model space M.

The principle of choosing a model that maximizes the likelihood function is well established in the "classical" school of statistics that is based on a frequency interpretation of probability. The justification for this principle here is quite different, because prior information is used explicitly. The model m^* is chosen after the experimental results E are known, and it is the strongest competing model[†] to m in the model space M. No other model will give a larger shift in the odds in going from the prior state of information (without E)

[†] Tribus [83], [8] refers to m^* as the "Monday morning quarterback hypothesis."

to the posterior state of information that includes E. Let us now examine how prior information about the models m and m* is used. Suppose that prior to knowing E the model m is considered much more probable than m*, say $ev(m{:}m^*|\mathcal{E}) = k$. (Of course, we did not know what m* was until after E was known, so we must make this assessment on the basis of the "past" state of information, $\mathcal{E}$). If the posterior probabilities assigned to m and to m* are to be roughly equivalent, the posterior evidence function must be approximately zero, implying that the logarithm of the likelihood ratio must be about -k. The log likelihood ratio must be negative, since m* maximized the likelihood function $\{E|m^*,\mathcal{E}\}$ over all models in the model space M, including the model m. If the log likelihood ratio is considerably greater than -k, it is clear that the posterior probability assigned to m is much greater than to the model m*. We can usually stop at this point: The data does not seem sufficient to overturn our belief in the model m. If the log likelihood is equal to or less than -k (i.e., more negative) on the other hand, the data plus prior information signify that our assumption that m is a suitable model of the process should be subject to serious question. The actual decision as to which model or models should be used for an analysis should include the economic considerations; the decreased value of the analysis if a simple invariance model is not appropriate must be balanced against the increased cost of a more complex model in which some of the invariance assumptions have been relaxed. These economic considerations may often be passed over quickly because the inferential aspects predominate: The log likelihood ratio either greatly exceeds or falls short of the threshold -k established by prior information.

Let us return to the coin tossing example. The fair coin model m implies that tosses are Benoulli trials with a probability of heads, p, equal to one-half. We are considering two possible experimental results, (a) a sequence containing nine heads in twenty tosses (see page 128), and (b) a sequence of twenty straight heads.

Suppose first that the model space M is the set of all possible sequences of heads and tails in twenty tosses. The model m^* specifies the sequence exactly, and there is no remaining uncertainty. If the experiment were to be repeated under conditions that we perceived to be identical then the same sequence will result from the second experiment as the first. The likelihood function $\{E|m^*,\mathcal{E}\} = 1$ for both sequences (a) and (b), and the likelihood ratio for m relative to m^* is $2^{-20} \cong 10^{-6}$, giving a log likelihood ratio of -60 decibels. Now let us examine the prior assignment to m^* for the case in which this model is based on the sequence (a). The model m^* implies that experiments consisting of twenty tosses of the coin generates exactly this sequence. Since there are 10^{-6} possible sequences of 20 tosses and since (a) has no readily apparent structure, it would seem reasonable to assign a very small prior probability to m^*: say 10^{-24}. This is equal to 10^{-6} of getting this particular sequence (all sequences considered equally probable on the basis of insufficient reason), times a very small probability (10^{-18}) that the sequence will repeat itself exactly on subsequent experiments. The prior odds are then -240 decibels, and the posterior odds -180 db, the prior probability assigned to sequences repeating, a very small number. At this point we doubt that the fair coin model should be rejected in favor of the hypothesis that every sequence of twenty tosses will produce the result (a); if sequence (a) resulted from four successive experiments, however, we would consider rejecting the fair coin model.

Suppose that the sequence (b), twenty straight heads, results. The model $m*$ is then that in every sequence of twenty tosses, all tosses will be heads. This hypothesis has more plausible explanations: A magician may have substituted a two-headed coin, or the coin is being tossed using a special coin-flipper that always results in the outcome heads.*

Therefore, in the case of (b) we might assign a prior to $m*$ that is much higher than for (a), say 10^{-6}. This assignment would imply that the posterior odds of m, compared to $m*$ are about even; we assign about the same posterior probability to the fair coin model as to the hypothesis that all tosses will be heads.

Suppose that we choose as our model space M the class of all possible exchangeable sequences. The tosses are presumed to be Bernoulli trials indexed by an unknown parameter p, the long-run fraction of heads = the probability of a head on any arbitrarily chosen toss. Let us consider the likelihood function $\{E|m*,\mathcal{E}\}$ for sequence (a). It is a maximum for $p = 0.45$. We find that the likelihood ratio for $p = 0.50$ (the fair coin model) versus $p = 0.45$ is about 0.90, and the log likelihood ratio is -0.4 db. The posterior evidence function is barely different from the prior evidence function; the posterior odds are reduced by a factor of ten per cent from the prior odds in

*Of course, such hypothesis might be checked directly. We shall pretend that such checks are not available.

this worst case. For any other model m_o chosen from the Bernoulli class, the difference between the prior odds for the fair coin models versus m_o and the posterior odds will be less.

For sequence (b) the likelihood function $\{E|m^*,\mathcal{E}\}$ attains a maximum for $p = 1$, and the log likelihood ratio is -60 db. The posterior odds for the fair coin versus a Bernoulli model with $p = 1$ are reduced by a factor of 10^6 from the prior odds.

For a sequence with two heads in twenty tosses we obtain a log likelihood ratio of -32 db. If the result had six heads in twenty tosses, we would get a log likelihood ratio of -7 db, corresponding to a reduction in the odds by a factor of about five.

We could also test the fair coin model against models in the Markov class, in which the probability of a head is allowed to depend on the result of the preceding toss. There would then be two parameters indexing the model space, and we would again select that model m^* that made $\{E|m^*,\mathcal{E}\}$ a maximum, and therefore the log likelihood ratio a minimum.

A key idea in taking m^* to be the worst alternative model is that we avoid the effort of encoding a prior probability distribution over the parameters indexing the model space M. If we are testing an invariance model m against a class of models M obtained by relaxing the invariance, it may be quite reasonable to assume that the model m^* in M that best fits the data (i.e., maximizes the likelihood function) will have a prior probability $\{m^*|\mathcal{E}\}$ that is representative of the class M. If the data would not lead to a large revision of the odds for m as compared to m^*, we have shown that the invariance model m is not inconsistent with the data. If data leads to substantial revision of the odds we may wish to question the assumed invariance and to examine in more detail the structure of our prior information regarding models in M. For example, we might have begun with the hypothesis that a process generating two outcomes, heads and tails, corresponds to a fair coin. We observe two heads in twenty tosses, and the large shift (-32 db) from prior to posterior odds comparing the fair coin hypothesis $(p = 0.5)$ to the maximum likelihood hypothesis $(p = 0.1)$ causes us to question the fair coin model. At this point we may wish to consider other possible models based on symmetry or invariance criteria (e.g., suppose the process is generated by rolling a fair die: a one corresponds to heads, two through six correspond to tails. Then $p = 1/6$), or to encode all of our relevant knowledge about the model class M by means of a probability distribution assigned to the set of indexing parameters (i.e., to p).

Let us now summarize our procedure once more for testing an invariance derived model m. From the expression (7.9), the logarithmic form of Bayes' Rule,

$$\mathrm{ev}(m:m^*|E,\mathcal{E}) - \mathrm{ev}(m:m^*|\mathcal{E}) = 10 \log_{10} \frac{\{E|m,\mathcal{E}\}}{\{E|m^*,\mathcal{E}\}} \qquad (7.10)$$

we find that the logarithm of the likelihood ratio for m relative to an alternative model m* gives a measure of the change in the odds we would assign to m versus m* based on prior and posterior information. The greatest shift in the odds occurs if m* is chosen to maximize the likelihood function; we shall use this "worst case" as a basis for testing the model m against the alternative models in a model space M. Providing m is in M, the log likelihood function will be non-positive; it will be zero only if m maximizes the likelihood function. We shall question or reject the model m only if the factor by which the odds are shifted (in going from prior to posterior information) exceeds a threshold, let us say k. If the shift in the odds is less than k, we shall retain the model m. Then if we define

$$\psi = 10 \log_{10} \frac{\{E|m,\mathcal{E}\}}{\{E|m^*,\mathcal{E}\}} \qquad (7.11)$$

the form of the test is as follows:

$0 \leq \psi < k$: Retain the model m

$\psi \geq k$: The model is declared to be in doubt, and further analysis is indicated.

The threshold k reflects our state of prior information, and possibly economic factors as well.

7.2 The Case of Exchangeable Sequences

Let us now specialize this test to the case of exchangeable sequences. Suppose that our experimental results E consist of n observations $x_1, \ldots, x_n$ from an exchangeable sequence. For the moment let us assume a discrete outcome space of N points, and let us compute the probability that out of the n observed results, n_1 will be the first outcome point, n_2 the second, and so forth through n_N for the N^{th} outcome point. The probability of the experimental result $E = x_1, \ldots, x_n$ is then

$$\{x_1, \ldots, x_n|m,\mathcal{E}\} = \frac{n!}{n_1!\, n_2! \cdots n_N!} p_1^{n_1} p_2^{n_2} \cdots p_N^{n_N} \qquad (7.12)$$

where p_i is the probability of the i^{th} outcome point assigned on the basis of the model m. If m* is the assumed model, a similar expression holds with probabilities $p'_1, \ldots , p'_N$. In computing the ratio of the likelihood functions the multinomial coefficients cancel. If we take ψ to be the (natural) logarithm of the ratio of the likelihood functions, we have

$$\begin{aligned} \psi &= \log\left(\frac{\{x_1, \ldots , x_n | m, \mathcal{E}\}}{\{x_1, \ldots , x_n | m*, \mathcal{E}\}}\right) \\ &= n \sum_{k=1}^{N} \left(\frac{n_k}{n}\right) \log\left(\frac{p_k}{p'_k}\right) . \end{aligned} \tag{7.13}$$

Now let the alternative model m* be that which maximizes the likelihood function $\{x_1, \ldots , x_N | m*, \mathcal{E}\}$ and therefore minimizes the likelihood ratio. This maximum is attained if p'_k is chosen to be the observed relative frequency of the k^{th} outcome:

$$p'_k = \frac{n_k}{n} = f_k \tag{7.14}$$

then

$$\psi = -n \sum_{k=1}^{N} (f_k \log f_k - f_k \log p_k) . \tag{7.15}$$

Recalling the inequality of (5.46),

$$- \sum_{k=1}^{N} f_k \log f_k \leq \sum_{k=1}^{N} f_k \log p_k \tag{7.16}$$

we see that $\psi \leq 0$, with equality only if $p_k = f_k$, as we might expect. The log likelihood ratio for exchangeable sequences resembles then the expression for the entropy, a result which we might expect on the basis of the combinational derivation.

Suppose that the relative frequencies f_k are close to the model predictions p_k. Then $(f_k - p_k) << 1$ and we can use the logarithmic expansion

$$\log_e(x) = (x-1) - \frac{1}{2}(x-1)^2 + \cdots . \tag{7.17}$$

Assuming a base e for the log likelihood ratio ψ, we have[*]

$$\begin{aligned} \psi &= n \sum_{k=1}^{N} f_k \log(p_k/f_k) \\ &= n \sum_{k=1}^{N} f_k \left[\left(\frac{p_k}{f_k} -1\right) - \frac{1}{2}\left(\frac{p_k}{f_k} - 1\right)^2 + \cdots\right] \end{aligned} \tag{7.18}$$

$$\psi \cong - \frac{1}{2} n \sum_{k=1}^{N} \frac{(p_k - f_k)^2}{f_k} . \tag{7.19}$$

If $p_k \cong f_k$, we can replace f_k in the denominator by p_k, and then -2ψ is exactly the well-known statistic developed by Karl Pearson to measure goodness of fit for a distribution. Since the quantity

$$\frac{f_k - p_k}{\sqrt{np_k(1 - p_k)}} \tag{7.20}$$

is approximately normally distributed with mean zero and variance one (assuming that the model m in fact holds true) for large n by the central limit theorem, the Pearson statistic

$$D^2 = n \sum_{k=1}^{N} \frac{(p_k - f_k)^2}{p_k} \tag{7.21}$$

is distributed asymptotically as a χ^2 random variable with N-1 degrees of freedom.

[*] Using 10 $\log_{10}$ as in the earlier section of this chapter simply introduces a multiplicative constant.

A much stronger result is available due to Wilks [92]. If the maximum likelihood estimates specifying the model m^* have an a priori distribution (that is, where E is unknown and is therefore considered a random variable) that is asymptotically normal, then the distribution of -2ψ is asymptotically χ^2 with r degrees of freedom, where r is the dimensionality of the model space M, and ψ is the logarithm (base e) of the likelihood ratio. This result is considered to be one of the fundamental achievements in the classical statistics approach to hypothesis testing (Mood and Graybill [57], Freeman [22], Wilks [92]).

From our point of view the use of the log likelihood ratio ψ directly for hypothesis testing seems like a more satisfying procedure. For any n, not just the large sample limit, the quantity ψ can be related to the shift from the prior to the posterior odds for m relative to the maximum likelihood (worst case) model m^*.

The use of the statistic ψ for testing distributions of independent, identically distributed random variables (i.e., equation (7.13)), and its relation to the classical χ^2 test developed by Karl Pearson has been noted by E. T. Jaynes [41]. Jaynes' development follows in part from the earlier work of I. J. Good [27].

Chapter VIII

SUMMARY AND CONCLUSION

To use probability theory for inductive reasoning about what is uncertain we must begin with something that is known, namely, an individual's state of information. We translate this state of information into probability assignments. If we are to use the elegant models of modern probability theory for inductive reasoning we need a conceptual basis to justify these models as a representation of the state of information.

The basic desideratum provides a means to achieve this conceptual basis. If two or more states of information are perceived as equivalent then the probabilistic model should be invariant to changes from one of these states to another. The principle of insufficient reason provides a simple illustration. If any relabeling of the possible outcomes leads to an equivalent state of information, then the probabilities assigned to the outcomes should remain invariant to changing these assignments from one set of outcomes to another, hence the probabilities assigned to each outcome must be equal.

The deterministic physical model of Gibbs' statistical mechanics provided another example of an invariance principle. For a system in equilibrium our state of information does not depend on time, hence the probability distribution assigned to initial conditions should not depend on the time at which these conditions are measured. We observed that this invariance principle led to a maximum entropy characterization of the probability distribution.

De Finetti's concept of exchangeability provides a means for applying invariance to a statistical ensemble, that is, a sequence of experiments performed under identical conditions. The state of information should be invariant to the sequence of the experiments, implying through the basic desideratum that joint probability assignments to experimental outcomes should be invariant to permutations in the order of the experiments. The concept of exchangeability may be generalized by considering the invariance to permutation to hold only for subsequences of experiments.

The inference problem for an exchangeable sequence is to characterize the histogram summarizing a large number of observed experimental outcomes. Since this histogram may have an unmanageably large number of degrees of freedom, it is often advisable to introduce an additional invariance principle, which we call the extended principle of insufficient reason. This principle asserts that if the state of information is characterized only by a set of constraints on the histogram, then all sequences of experimental outcomes satisfying these constraints should be viewed as equally probable. By computing the number of sequences resulting in a given histogram we are led to an entropy measure for the probability of a given histogram; the maximum entropy principle is equivalent to choosing the histogram that can be realized by the largest number of (equally probable) sequences. When the constraints take the form that an average value defined on the sequence takes on a specific (but uncertain) value, the probability distributions derived from the extended principle of insufficient reason are characterized by sufficient statistics for the inference of this value. Many of the common distributions of probability theory, such as the normal, exponential, and gamma distributions, possess this characterization.

The equivalence between states of information that forms the basis for the invariance approach is a situation connoting comparative ignorance. With further information gained from additional data we may decide that the original equivalence in states of information was unwarranted, and a more complex encoding procedure should be used. The test of invariance-derived models by means of Bayes' Rule is complicated by the dimensionality of the space of alternative models. We avoid this dimensionality by taking a "worst case" approach and using Bayes' Rule in a logarithmic form. The resulting model test is asymptotically equivalent to the likelihood ratio tests of classical statistics and Pearson's traditional χ^2 test.

The invariance approach that has been explored in this dissertation provides a means by which probability assignments and probabilistic models may be related to a state of information. We have discussed a number of successful applications of the invariance approach. Although we cannot assert that this method provides a comprehensive solution to the probabilitic encoding of information, it appears to offer a conceptual unity that Bayesian probability has heretofore lacked. Much additional effort will be needed before the full potential of the invariance approach will be apparent. It is hoped that this approach will prove to be a significant step toward a unified methodology of inductive reasoning.

REFERENCES

1. Abramson, Norman, Information Theory and Coding, McGraw-Hill, New York, (1963).

2. Aoki, Masanao, Optimization of Stochastic Systems, Academic Press, New York, (1967).

3. Bather, J. A., "Bayes Procedures for Deciding the Sign of a Normal Mean," Proceedings of the Cambridge Philosophical Society, Vol. 58, 599-620, (1962).

4. Bather, John and Herman Chernoff, "Sequential Decisions in the Control of a Spaceship," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L. LeCam and J. Neyman, eds., University of California Press, Berkeley, California, Vol. III, 181-207, (1967).

5. Brewer, K. R. W., "Decisions under Uncertainty: Comment," Quarterly Journal of Economics, Vol. 77, No. 1, 159-161, (1963).

6. Chernoff, Herman, "Rational Selection of Decision Functions," Econometrica, Vol. 22, No. 4, 422-443, (1954).

7. Chernoff, Herman, "Sequential Tests for the Mean of a Normal Distribution," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., Vol. I, 79-91, (1960).

8. Chernoff, Herman, and S. N. Ray, "A Bayes Sequential Sampling Inspection Plan," Annals of Mathematical Statistics, Vol. 36, 1387-1407, (1965).

9. Cox, Richard T., The Algebra of Probable Inference, The Johns Hopkins Press, Baltimore, Maryland, (1961).

10. Davenport, Wilbur B., Jr., and William L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, New York, (1959).

11. DeGroot, M. H., "Uncertainty, Information, and Sequential Experiments," Annals of Mathematical Statistics, Vol. 33, No. 2, 404-419, (1962).

12. Doob, J. L., Stochastic Processes, Wiley, New York, (1953).

13. Edwards, Ward, Lawrence D. Phillips, William L. Hays, and Barbara C. Goodman, "Probabilistic Information Processing Systems: Design and Evaluation," IEEE Transaction on Systems Science and Cybernetics, Vol. SSC-4, No. 3, 248-265, (1968).

14. Ehrenberg, W., "Maxwell's Demon," Scientific American, Vol. 217, No. 5, 103-110, (1967).

15. Ellsberg, Daniel, "Risk, Ambiguity, and the Savage Axioms," Quarterly Journal of Economics, Vol. 75, 643-669, (1961).

16. Feinstein, Amiel, Foundations of Information Theory, McGraw-Hill, New York, (1958).

17. Feller, William, An Introduction to Probability Theory and Its Applications, Vol. I, second edition, Wiley, New York, (1957).

18. Fellner, William, "Distortion of Subjective Probabilities as a Reaction to Uncertainty," Quarterly Journal of Economics, Vol. 75, 670-689, (1961).

19. Fellner, William, "Slanted Subjective Probabilities and Randomization: Reply to Howard Raiffa and K. R. W. Brewer," Quarterly Journal of Economics, Vol. 77, No. 4, 676-690, (1963).

20. Finetti, Bruno de, "La prévision: ses lois logiques, ses sources subjectives," Annals de l'Institut Henri Poincaré, Vol. 7, 1-68, (1937). Trans. by Henry E. Kyburg, Jr. as "Foresight: Its Logical Laws, Its Subjective Sources," Studies in Subjective Probability, Kyberg and Smokler, ed., Wiley, New York, 97-158, (1963).

21. Finetti, Bruno de, "Sur la condition d'équivalence partielle," Actualities Scientifiques et Industrielles, Vol. 739, 5-18, (1938).

22. Freeman, Harold, Introduction to Statistical Inference, Addison-Wesley, Reading, Massachusetts, (1963).

23. Gibbs, J. Willard, Elementary Principles in Statistical Mechanics, Yale University Press, New Haven, Connecticut, (1902). Reprinted by Dover Publications, New York, 1960.

24. Girschick, M. A., and Herman Rubin, "A Bayes Approach to a Quality Control Model," The Annals of Mathematical Statistics, Vol. 23, 114-125, (1952).

25. Gnedenko, B. V., and A. N. Kolmogorov, Limit Distributions for Sums of Independent Random Variables, Addison-Wesley, Cambridge, Massachusetts, (1954).

26. Goldstein, Herbert, Classical Mechanics, Addison-Wesley, Redding, Massachusetts, (1959).

27. Good, I. J., Probability and the Weighing of Evidence, Griffin, London, (1950).

28. Good, I. J., "Maximum Entropy for Hypothesis Formulation," The Annals of Mathematical Statistics, Vol. 34, 911-930, (1963).

29. Howard, Ronald A., "Information Value Theory," IEEE Transactions on System Science and Cybernetics, Vol. SSC-2, No. 1, 22-26, (1966).

30. Howard, Ronald A., "The Foundations of Decision Analysis," IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-4, No. 3, 211-219, (1968).

31. Huang, Kerson, Statistical Mechanics, Wiley, New York, (1963).

32. Hurwicz, Leonid, "Programming in Linear Spaces," Studies in Linear and Non-Linear Programming, by Arrow, Hurwicz, and Uzawa, Stanford Univ. Press, Stanford, California, 38-102, (1958).

33. Jaynes, Edwin T., "Information Theory and Statistical Mechanics," The Physical Review, Vol. 106, No. 4, 620-630, (1957).

34. Jaynes, Edwin T., "Information Theory and Statistical Mechanics II," The Physical Review, Vol. 108, No. 2, 171-190, (1957).

35. Jaynes, Edwin T., Probability Theory in Science and Engineering, Field Research Laboratory, Socony Mobil Oil Company, Inc., Dallas, Texas, (1959).

36. Jaynes, Edwin T., "Information Theory and Statistical Mechanics," (1962 Brandeis Lectures), Chapter 4 of Statistical Physics, K. W. Ford, ed., W. A. Benjamin, Inc., New York, 182-218, (1963).

37. Jaynes, Edwin T., "New Engineering Applications of Information Theory," Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability, Wiley, New York, 163-203, (1963).

38. Jaynes, Edwin T., "Gibbs vs. Boltzmann Entropies," American Journal of Physics, Vol. 33, No. 5, (1965).

39. Jaynes, Edwin T., "Foundations of Probability Theory and Statistical Mechanics," Chapter 6 of *Delaware Seminar in the Foundations of Physics*, M. Bunge, ed., Springer, Berlin, (1967).

40. Jaynes, Edwin T., "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics*, Vol. SSC-4, No. 3, 227-241, (1968).

41. Jaynes, Edwin T., *Probability Theory in Science and Engineering*, A series of ten lectures: mimeographed notes (1960). This is an expanded version of reference [35], which contains only the first five lectures.

42. Jeffreys, Harold, "An Extension of the Pitman-Koopman Theorem," *Proceedings of the Cambridge Philosophical Society*, Vol. 56, Part 4, 393-395, (1960).

43. Jeffreys, Harold, *Theory of Probability*, (first edition, 1939), third edition, the Clarendon Press, Oxford, (1961).

44. Keynes, John Maynard, *A Treatise on Probability*, MacMillan, London, (1921).

45. Khinchin, A. I., *Mathematical Foundations of Information Theory*, trans. by R. A. Silverman and M. D. Friedman, Dover, New York, (1957).

46. Koopman, B. O., "On Distributions Admitting a Sufficient Statistic," *Transactions of the American Mathematical Society*, Vol. 39, 399-409, (1936).

47. Laplace, Pierre Simon, Marquis de, *Essai philosophique sur les Probabilités*, Paris, (1814); *A Philosophical Essay on Probabilities*, (tr.), Dover Publications, New York, (1951).

48. Lindley, D. V., "Dynamic Programming and Decision Theory," Applied Statistics, Vol. 10, No. 1, 39-51, (1961).

49. Lindley, D. V., Introduction to Probability and Statistics from a Bayesian Viewpoint Part 1: Probability, Part 2: Inference, Cambridge University Press, (1965).

50. Lindley, D. V., and B. N. Barnett, "Sequential Sampling: Two Decision Problems with Linear Losses for Binomial and Normal Random Variables," Biometrika, Vol. 52, No. 3-4, 507-532, (1965).

51. Loève, Michel, Probability Theory, D. Van Nostrand, Princeton, New Jersey, (1962).

52. Luce, R. Duncan and Howard Raiffa, Games and Decisions: Introduction and Critical Survey, Wiley, New York, (1957).

53. Luenberger, David G., Optimization by Vector Space Methods, Wiley, New York, (1968).

54. Marschak, Jacob, and Koichi Miyasawa, "Economic Comparability of Information Systems," International Economic Review, Vol. 9, No. 2, 137-174, (1968).

55. Matheson, James E., "The Economic Value of Analysis and Computation," IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-4, No. 3, 325-332, (1968).

56. Matheson, James E., and Arnold B. Pollard, "A New Approach to Space Project Planning: Decision Analysis of Voyager," Final Report Project 6152, Stanford Research Institute, Menlo Park, California, (1967).

57. Mood, Alexander M., and Franklin A. Graybill, Introduction to the Theory of Statistics, second edition, McGraw-Hill, New York, (1963).

58. Murray, George R., Jr., "Modeling and the Use of Subjective Knowledge in Engineering Design and Experiment Selection," 1966 Conference Record, IEEE Systems Science and Cybernetics Conference, Washington, D. C., (1966).

59. Murray, George R., and Richard D. Smallwood, "Model Inference in Adaptive Pattern Recognition," Proceedings of the Fourth International Conference of the International Federation of Operations Research Societies, Wiley, New York, 92-110, (1966).

60. North, D. Warner, "A Tutorial Introduction to Decision Theory," IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-4, No. 3, 200-210, (1968).

61. Parzen, Emanuel, Modern Probability Theory and Its Applications, Wiley, New York, (1960).

62. Parzen, Emanuel, Stochastic Processes, Holden-Day, San Francisco, (1962).

63. Pitman, E. J. G., "Sufficient Statistics and Intrinsic Accuracy," Proceedings of the Cambridge Philosophical Society, Vol. 32, 567-579, (1936).

64. Pratt, John W., "The Outer Needle of Some Bayes Sequential Continuation Regions," Biometrika, Vol. 53, No. 3-4, 455-467, (1966).

65. Pratt, John W., Howard Raiffa, and Robert Schlaifer, "The Foundations of Decision Under Uncertainty: An Elementary Exposition," Journal of the American Statistical Association, Vol. 59, No. 306, 353-375, (1964).

66. Pratt, John W., Howard Raiffa, and Robert Schalifer, Introduction to Statistical Decision Theory, McGraw-Hill, New York, (1965).

67. Quisel, Kent, "Extensions of the Two-Armed Bandit and Related Processes with On-Line Experimentation," Technical Report No. 137, Institute for Mathematical Studies in the Social Sciences, Stanford University, (1965).

68. Raiffa, Howard, Decision Analysis: Introductory Lectures on Choices under Uncertainty, Addison-Wesley, Reading, Massachusetts, (1968).

69. Raiffa, Howard, Unpublished data, presented at the decision analysis seminar at the Graduate School of Business, Stanford University, May, (1969).

70. Raiffa, Howard, and Robert Schlaifer, Applied Statistical Decision Theory, Graduate School of Business, Harvard University, Boston, Massachusetts, (1961).

71. Raiffa, Howard, "Risk, Ambiguity, and the Savage Axioms: Comment," Quarterly Journal of Economics, Vol. 75, 690-694, (1961).

72. Ramsey, Frank P., "Truth and Probability," The Foundations of Mathematics and Other Logical Essays, R. B. Braithwaite, ed., Humanities Press, New York, (1950). Reprinted in Studies in Subjective Probability, Kyberg and Smokler, eds., Wiley, New York, 63-92, (1963).

73. Roberts, Harry V., "Risk, Ambiguity, and the Savage Axioms: Comment," Quarterly Journal of Economics, Vol. 77, No. 2, 327-342, (1963).

74. Savage, Leonard J., The Foundations of Statistics, Wiley, New York, (1954).

75. Savage, Leonard J., and others, The Foundations of Statistical Inference, Methuen, London, (1962).

76. Scarf, Herbert, "Bayes Solutions of the Statistical Inventory Problem," Annals of Mathematical Statistics, Vol. 30, No. 2, 490-508, (1959).

77. Schrődinger, Erwin, Statistical Thermodynamics, (Dublin lectures, 1944), Cambridge University Press, (1952).

78. Shannon, Claude E., "The Mathematical Theory of Communication," Bell System Technical Journal, Vol. 27, 379, 623, July and October, (1948). Reprinted in C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, Illinois, (1949).

79. Smallwood, Richard D., "A Decision Analysis of Model Selection," IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-4, No. 3, 333-342, (1968).

80. Smith, Vernon L., "Measuring Nonmonetary Utilities in Uncertain Choices: the Ellsberg Urn," Quarterly Journal of Economics, No. 331, Vol. 83, 324-329, (1969).

81. Spragins, J. D., Jr., Reproducing Distributions for Machine Learning, Stanford Electronics Laboratories, Technical Report No. 6103-7, (1963).

82. Tribus, Myron, Thermostatics and Thermodynamics, D. Van Nostrand Co., Princeton, New Jersey, (1961).

83. Tribus, Myron, "The Use of the Maximum Entropy Estimate in Reliability," Recent Developments in Information and Decision Processes, Robert E. Machol and Paul Gray, ed., MacMillan, New York, (1962).