FINITE MARKOV CHAINS

by

JOHN G. KEMENY
Professor of Mathematics
Dartmouth College

AND

J. LAURIE SNELL
Associate Professor of Mathematics
Dartmouth College

D. VAN NOSTRAND COMPANY, INC.
PRINCETON, NEW JERSEY
TORONTO LONDON
NEW YORK

D. VAN NOSTRAND COMPANY, INC.
120 Alexander St., Princeton, New Jersey (*Principal office*)
257 Fourth Avenue, New York 10, New York

D. VAN NOSTRAND COMPANY, LTD.
358, Kensington High Street, London, W.14, England

D. VAN NOSTRAND COMPANY (Canada), LTD.
25 Hollinger Road, Toronto 16, Canada

Published simultaneously in Canada by
D. VAN NOSTRAND COMPANY (Canada), LTD.

Library of Congress Catalogue Card No. 59-15644

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

The basic concepts of Markov chains were introduced by A. A. Markov in 1907. Since that time Markov chain theory has been developed by a number of leading mathematicians. It is only in very recent times that the importance of Markov chain theory to the social and biological sciences has become recognized. This new interest has, we believe, produced a real need for a treatment, in English, of the basic ideas of finite Markov chains.

By restricting our attention to finite chains, we are able to give quite a complete treatment and in such a way that a minimum amount of mathematical background is needed. For example, we have written the book in such a way that it can be used in an undergraduate probability course, as well as a reference book for workers in fields outside of mathematics.

The restriction of this book to finite chains has made it possible to give simple, closed-form matrix expressions for many quantities usually given as series. It is shown that it suffices for all types of problems to consider just two types of Markov chains, namely absorbing and ergodic chains. A "fundamental matrix" is developed for each type of chain, and the other interesting quantities are obtained from the fundamental matrices by elementary matrix operations.

One of the practical advantages of this new treatment of the subject is that these elementary matrix operations can easily be programed for a high-speed computer. The authors have developed a pair of programs for the IBM 704, one for each type of chain, which will find a number of interesting quantities for a given process directly from the transition matrix. These programs were invaluable in the computation of examples and in the checking of conjectures for theorems.

A significant feature of the new approach is that it makes no use of the theory of eigen-values. The authors found, in each case, that the expressions in matrix form are simpler than the corresponding expressions usually given in terms of eigen-values. This is presumably due to the fact that the fundamental matrices have direct probabilistic interpretations, while the eigen-values do not.

The book falls into three parts. Chapter I is a very brief summary of prerequisites. Chapters II–VI develop the theory of Markov chains. Chapter VII contains applications of this theory to problems in a variety of fields. A summary of the symbols used and of the principal definitions and formulas can be found in the appendices together with page references. Therefore, there is no index, but it is hoped that the de-

tailed table of contents and the appendices will serve a more useful purpose.

It was not intended that Chapter I be read as a unit. The book can be started in Chapter II, and the reader has the option of looking up the brief summary of any prerequisite topic not familiar to him, when he needs it in a later chapter.*

The book was designed so that it can be used as a text for an undergraduate mathematics course. For this reason the proofs were carried out by the most elementary methods possible. The book is suitable for a one-semester course in Markov chains and their applications. Selections from the book (presumably from Chapters II, III, IV, and possibly VII) could also be used as part of an upper-class course in probability theory. For this use, exercises have been given at the end of Chapters II–VI.

The following system of notation has been used in the book: Numbers are denoted by small italic letters, matrices by capital italics, vectors by Greek letters. Functions, sets, and other abstract objects are denoted by boldface letters.

The authors gratefully acknowledge support by the National Science Foundation to the Dartmouth Mathematics Project. Many of the original results in this book were found by the authors while working on this project. The authors are also grateful for computing time made available by the M.I.T. and Dartmouth Computation Center for the development of the above-mentioned programs and for the use of these programs.

The authors wish to express their thanks to two research assistants, P. Perkins and B. Barnes, for many valuable suggestions as well as for their careful reading of the manuscript. Thanks are due to Mrs. M. Andrews and Mrs. H. Hanchett for typing the manuscript.

THE AUTHORS

Hanover, New Hampshire
September, 1959

* A more detailed treatment of most of these topics may be found in one of the following books: (1) *Modern Mathematical Methods and Models,* Volumes 1 and 2, by the Dartmouth Writing Group, published by the Mathematical Association of America, 1958. [Referred to as M[4].] (2) *Introduction to Finite Mathematics,* by Kemeny, Snell, and Thompson, Prentice-Hall, 1957. [Referred to as FM.] (3) *Finite Mathematical Structures,* by Kemeny, Mirkil, Snell, and Thompson, Prentice-Hall, 1959. [Referred to as FMS.] For the prerequisites in probability theory, as well as a treatment of Markov chains from a different point of view, the reader may also wish to consult *Introduction to Probability Theory and Its Applications,* by W. Feller, Wiley, 1957.

TABLE OF CONTENTS

CHAPTER I—PREREQUISITES

CHAPTER II—BASIC CONCEPTS OF MARKOV CHAINS

CHAPTER III—ABSORBING MARKOV CHAINS

CHAPTER IV—REGULAR MARKOV CHAINS

CHAPTER I

PREREQUISITES

§ 1.1 Sets. By a *set* a mathematician means an arbitrary but well-defined collection of objects. Sets will be denoted by bold capital letters. The objects in the collection are called *elements*.

If $\mathbf{A}$ is a set, and $\mathbf{B}$ is a set whose elements are some (but not necessarily all) of the elements of $\mathbf{A}$, then we say that $\mathbf{B}$ is a *subset* of $\mathbf{A}$, symbolized as $\mathbf{B} \subseteq \mathbf{A}$. If the two sets have exactly the same elements, then we say that they are equal, i.e. $\mathbf{A} = \mathbf{B}$. Thus $\mathbf{A} = \mathbf{B}$ if and only if $\mathbf{A} \subseteq \mathbf{B}$ and $\mathbf{B} \subseteq \mathbf{A}$. If $\mathbf{B}$ is a subset of $\mathbf{A}$ and is not equal to $\mathbf{A}$, then we say that it is a *proper subset*, and write $\mathbf{B} \subset \mathbf{A}$. If $\mathbf{A}$ and $\mathbf{B}$ have no element in common, we say that they are *disjoint*.

Very frequently we will deal with a given set of objects, and discuss various subsets of it. The entire set will be called the *universe*, $\mathbf{U}$. A particularly interesting subset is the set with no elements, the *empty set* $\mathbf{E}$.

Given a set, there are a number of ways of getting new subsets from old ones. If $\mathbf{A}$ and $\mathbf{B}$ are both subsets of $\mathbf{U}$, then we define the following operations:

(1) The *complement* of $\mathbf{A}$, $\tilde{\mathbf{A}}$, has as elements all the elements of $\mathbf{U}$ which are not in $\mathbf{A}$.

(2) The *union* of $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \cup \mathbf{B}$, has as elements all the elements of $\mathbf{A}$ and all the elements of $\mathbf{B}$.

(3) The *intersection* of $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \cap \mathbf{B}$, has as elements all the elements that $\mathbf{A}$ and $\mathbf{B}$ have in common.

(4) The difference of $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} - \mathbf{B}$ has as elements all the elements of $\mathbf{A}$ that are not in $\mathbf{B}$.

To illustrate these operations, we will list some easily provable relations between these sets:

$$\tilde{\mathbf{U}} = \mathbf{E} \qquad \widetilde{\mathbf{A} \cup \mathbf{B}} = \tilde{\mathbf{A}} \cap \tilde{\mathbf{B}} \qquad \mathbf{A} \cap \mathbf{B} = \mathbf{B} \cap \mathbf{A}$$

$$\tilde{\tilde{\mathbf{A}}} = \mathbf{A} \qquad \widetilde{\mathbf{A} \cap \mathbf{B}} = \tilde{\mathbf{A}} \cup \tilde{\mathbf{B}} \qquad \mathbf{A} \cup \mathbf{E} = \mathbf{A}$$

$$\mathbf{A} - \mathbf{B} = \mathbf{A} \cap \tilde{\mathbf{B}} \qquad \mathbf{A} \cup \mathbf{B} = \mathbf{B} \cup \mathbf{A} \qquad \mathbf{A} \cap \mathbf{E} = \mathbf{E}$$

If $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_r$ are subsets of $\mathbf{U}$, and every element of $\mathbf{U}$ is in one and only one set $\mathbf{A}_j$, then we say that $\mathbf{A}=\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_r\}$ is a *partition* of $\mathbf{U}$.

If we wish to specify a set by listing its elements, we write the elements inside curly brackets. Thus, for example, the set of the first five positive integers is $\{1, 2, 3, 4, 5\}$. The set $\{1, 3, 5\}$ is a proper subset of it. The set $\{2\}$, which is also a subset of the five-element set, is called a *unit set*, since it has only one element.

In the course of this book we will have to deal with both finite and infinite sets, i.e. with sets having a finite number or an infinite number of elements. The only infinite sets that are used repeatedly are the set of integers $\{1, 2, 3, \ldots\}$ and certain simple subsets of this set.

For a more detailed account of the theory of sets see FM Chapter II or FMS Chapter II.†

§ 1.2 Statements. We are concerned with a process which will frequently be a scientific experiment or a game of chance. There are a number of different possible outcomes, and we will consider various statements about the outcome.

We form the set $\mathbf{U}$ of all logically possible outcomes. These must be so chosen that we are assured that exactly one of these will take place. The set $\mathbf{U}$ is called the *possibility space*. If $\mathbf{p}$ is any statement about the outcome, then it will (in general) be true according to some possibilities, and false according to others. The set $\mathbf{P}$ of all possibilities which would make $\mathbf{p}$ true is called the *truth set* of $\mathbf{p}$. Thus to each statement about the outcome we assign a subset of $\mathbf{U}$ as a truth set. The choice of $\mathbf{U}$ for a given experiment is not unique. For example, for two tosses of a coin we may analyze the possibilities as $\mathbf{U}=\{\mathbf{HH}, \mathbf{HT}, \mathbf{TH}, \mathbf{TT}\}$ or $\mathbf{U}=\{0\mathbf{H}, 1\mathbf{H}, 2\mathbf{H}\}$. In the first case we give the outcome of each toss and in the second only the number of heads which turn up. (For a more detailed discussion of this concept see FM Chapter II or FMS Chapter II.)

Given two statements $\mathbf{p}$ and $\mathbf{q}$ having the same subject matter (i.e. the same $\mathbf{U}$), we have a number of ways of forming new statements from them. (We will assume that the statements have $\mathbf{P}$ and $\mathbf{Q}$ as truth sets:)

(1) The statement $\sim\mathbf{p}$ (read "not $\mathbf{p}$") is true if and only if $\mathbf{p}$ is false. Hence it has $\tilde{\mathbf{P}}$ as truth set.

(2) The statement $\mathbf{p}\vee\mathbf{q}$ (read "$\mathbf{p}$ or $\mathbf{q}$") is true if either $\mathbf{p}$ is true or $\mathbf{q}$ is true or both. Hence it has $\mathbf{P}\cup\mathbf{Q}$ as truth set.

† FM = Kemeny, Snell, and Thompson, *Introduction to Finite Mathematics*, Englewood Cliffs, N.J., Prentice-Hall, Inc., 1957.

FMS = Kemeny, Mirkil, Snell, and Thompson, *Finite Mathematical Structures*, Englewood Cliffs, N.J., Prentice-Hall, Inc., 1959.

(3) The statement $\mathbf{p} \wedge \mathbf{q}$ (read "**p** and **q**") is true if both **p** and **q** are true. Hence it has $\mathbf{P} \cap \mathbf{Q}$ as truth set.

Two special kinds of statements are among the principal concerns of logic. A statement that is true for each logically possible outcome, that is, a statement having **U** as its truth set, is said to be *logically true* (such a statement is sometimes called a tautology). A statement that is false for each logically possible outcome, that is a statement having **E** as its truth set, is *logically false* or *self-contradictory*.

Two statements are said to be *equivalent* if they have the same truth set. That means that one is true if and only if the other is true.

The statements $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k$ are *inconsistent* if the intersection of their truth sets is empty, i.e., $\mathbf{P}_1 \cap \mathbf{P}_2 \cap \cdots \cap \mathbf{P}_k = \mathbf{E}$. Otherwise they are said to be *consistent*. If the statements are inconsistent, then they cannot all be true. If they are consistent, then they could all be true.

The statements $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k$ are said to form a *complete set of alternatives* if for every element of **U** exactly one of them is true. This means that the intersection of any two truth sets is empty, and the union of all the truth sets is **U**. Thus the truth sets of a complete set of alternatives form a partition of **U**. A complete set of alternatives provides a new way (and normally a less detailed way) of analyzing the possible outcomes.

§ **1.3 Order relations.** We will need some simple ideas from the theory of order relations. A complete treatment of this theory will be found in M[4], Vol. II, Unit 2.† We will take only a few concepts from that treatment.

Let **R** be a relation between two objects (selected from a specified set **U**). We denote by **aRb** the fact that **a** holds the relation **R** to **b**. Some special properties of such relations are of interest to us.

1.3.1 DEFINITION. *The relation* **R** *is* reflexive *if* **xRx** *holds for all* **x** *in* **U**.

1.3.2 DEFINITION. *The relation* **R** *is* symmetric *if whenever* **xRy** *holds, then* **yRx** *also holds, for all* **x**, **y** *in* **U**.

1.3.3 DEFINITION. *The relation* **R** *is* transitive *if whenever* $\mathbf{xRy} \wedge \mathbf{yRz}$ *holds, then* **xRz** *also holds, for all* **x**, **y**, **z** *in* **U**.

1.3.4 DEFINITION. *A relation that is reflexive, symmetric, and transitive is an* equivalence relation.

The fundamental property of an equivalence relation is that it partitions the set **U**. More specifically, let us suppose that **R** is an

† M[4] = *Modern Mathematical Methods and Models*, by the Dartmouth Writing Group. Mathematical Association of America, 1958.

equivalence relation defined on **U**. We put elements of **U** into classes in such a manner that two elements **a** and **b** are in the same class if **aRb**. It can be shown that the resulting classes are well defined and mutually exclusive, giving us a partition of **U**. These classes are the *equivalence classes* of **R**.

For example, let **xRy** express that "**x** is the same height as **y**," where **U** is a set of human beings. Then the resulting partition divides these people according to their heights. Two men are in the same equivalence class if and only if they are the same height.

1.3.5 DEFINITION. *A relation* **T** *is said to be* consistent *with the equivalence relation* **R** *if, given that* **xRy**, *then if* **xTz** *holds so does* **yTz**, *and if* **zTx** *holds so does* **zTy**.

1.3.6 DEFINITION. *A relation that is reflexive and transitive is known as a* weak ordering relation.

A weak ordering relation can be used to order the elements of **U**. Given a weak ordering **T**, and given any two elements **a** and **b** of **U**, there are four possibilities: (1) $\mathbf{aTb} \wedge \mathbf{bTa}$; then the two elements are "alike" according to **T**. (2) $\mathbf{aTb} \wedge \sim(\mathbf{bTa})$; then **a** is "ahead" of **b**. (3) $\sim(\mathbf{aTb}) \wedge \mathbf{bTa}$; then **b** is "ahead." (4) $\sim(\mathbf{aTb}) \wedge \sim(\mathbf{bTa})$; then we are unable to compare the two objects.

For example, if **xTy** expresses that "I like **x** at least as well as **y**," then the four cases correspond to "I like them equally," "I prefer **x**," "I prefer **y**," and "I cannot choose," respectively.

The relation of being alike acts as an equivalence relation. Indeed, it can be shown that if **T** is a weak ordering, then the relation **xRy** that expresses that $\mathbf{xTy} \wedge \mathbf{yTx}$ is an equivalence relation consistent with **T**. Thus **T** serves both to classify and to order. Consistency assures us that equivalent elements of **U** have the same place in the ordering.

For example, if we choose "is at least as tall" as our weak ordering, this determines the equivalence relation "is the same height," which is consistent with the original relation.

1.3.7 DEFINITION. *If* **T** *is a weak ordering, then the relation* $\mathbf{xTy} \wedge \mathbf{yTx}$ *is the equivalence relation* determined *by it.*

1.3.8 DEFINITION. *If* **T** *is a weak ordering, and the equivalence relation determined by it is the identity relation* $(\mathbf{x}=\mathbf{y})$ *then* **T** *is a* partial ordering.

The significance of a partial ordering is that no two distinct elements are alike according to it. One simple way of getting a partial ordering is as follows: Let **T** be a weak ordering defined on **U**. Define a new relation **T*** on the set of equivalence classes by saying that **uT*v** holds if every element of **u** bears the relation **T** to every element of **v**.

This is a partial ordering of the equivalence classes, and we call it the partial ordering *induced* by **T**.

1.3.9 DEFINITION. *An element* **a** *of* **U** *is called a* minimal element *if* **aTx** *implies* **xTa** *for all* $\mathbf{x} \in \mathbf{U}$. *If a minimal element is unique, we call it a* minimum.

We can define "maximal element" and "maximum" similarly. If **U** is a finite set, then it is easily shown that for any weak ordering there must be at least one minimal element. However, this minimal element need not be unique. Similarly, the weak ordering must have a maximal element, but not necessarily a maximum.

§ 1.4 Communication relations. An important application of order relations is the study of communication networks. Let us suppose that **r** individuals are connected through a complex network. Each individual can pass a message on to a subset of the individuals. This we will call *direct contact*. These messages may be relayed, and relayed again, etc. This will be *indirect contact*. It will *not* be assumed that a member can contact himself directly. Let **aTb** express that the individual **a** can contact **b** (directly or indirectly) or that $\mathbf{a} = \mathbf{b}$. It is easy to verify that **T** is a weak ordering of the set of individuals. It determines the equivalence relation $\mathbf{xTy} \wedge \mathbf{yTx}$, which may be read as "**x** and **y** can communicate with each other, or $\mathbf{x} = \mathbf{y}$."

This equivalence relation may be used to classify the individuals. Two men will be in the same equivalence class if they can communicate, that is, if each can contact the other one. The induced partial ordering **T*** has a very intuitive meaning: The relation **uT*v** holds if all members of the class **u** can contact all members of the class **v**, but not conversely unless $\mathbf{u} = \mathbf{v}$. Thus the partial ordering shows us the possible flow of information.

In particular, **u** is a maximal element of the partial ordering if its members cannot be contacted by members of any other class, and **u** is a minimal element if its members cannot contact members of other classes. Thus the maximal sets are message initiators, while the minimal sets are terminals for messages. (See M[4] Vol. II, Unit 2.)

It is interesting to study a given equivalence class. Any two members of such a class can communicate with each other. Hence any member can contact any other member. But how long does it take to contact other members? As a unit of time we will take the time needed to send a message from any one member to any member he can contact directly. We call this one *step*. We will assume that member i sends out a message, and we will be interested to know where the message could possibly be after n steps.

Let $\mathbf{N}_{ij}$ be the set of n such that a message starting from member i can be in member j's hands at the end of n steps. We will first consider $\mathbf{N}_{ii}$, the possible times at which a message can return to its originator. It is clear that if $a \in \mathbf{N}_{ii}$ and $b \in \mathbf{N}_{ii}$, then $a+b \in \mathbf{N}_{ii}$ after all the message can return in a steps and can be sent out again and be received back after b more steps. So the set $\mathbf{N}_{ii}$ is closed under addition. The following number-theoretic result will be useful. Its proof is given at the end of the section.

1.4.1 THEOREM. *A set of positive integers that is closed under addition contains all but a finite number of multiples of its greatest common divisor.*

If the greatest common divisor of the elements of $\mathbf{N}_{ii}$ is designated d_i, it is clear that the elements of $\mathbf{N}_{ii}$ are all multiples of d_i. But Theorem 1.4.1 tells us in addition that all sufficiently high multiples of d_i are in the set.

Since each member can contact every other member in its equivalence class, the $\mathbf{N}_{ij}$ are non-empty. We next prove that for i and j in the same equivalence class, $d_i = d_j = d$, and that the elements of a given $\mathbf{N}_{ij}$ are congruent to each other modulo d (their difference is a multiple of d). Suppose that $a \in \mathbf{N}_{ij}$, $b \in \mathbf{N}_{ij}$, and $c \in \mathbf{N}_{ji}$.

First of all, member i can contact himself by sending a message to member j and getting a message back. Hence $a+c \in \mathbf{N}_{ii}$. The message could also go to member j, come back to member j, and then go to member i. This could be done in $a+kd_j+c$ steps, where k is sufficiently large. Hence d_j must be a multiple of d_i. But in exactly the same way we can prove that d_i is a multiple of d_j. Hence $d_i = d_j = d$.

Or again, the message could go to member j in b steps, and then back to member i. Hence $b+c \in \mathbf{N}_{ii}$. Hence $a+c$ and $b+c$ are both divisible by d, and thus we see that $a \equiv b \pmod{d}$. Thus the elements of a given $\mathbf{N}_{ij}$ are congruent to each other modulo d. We can thus introduce numbers t_{ij}, with $0 \leqslant t_{ij} < d$, so that any element of $\mathbf{N}_{ij}$ is congruent to t_{ij}, modulo d. It is also easy to see that $\mathbf{N}_{ij}$ contains all but a finite number of the numbers $t_{ij}+kd$.

In particular we see that $t_{ii} = 0$ in each case, and hence $t_{ij}+t_{ji} \equiv 0 \pmod{d}$. Also $t_{ij}+t_{jm} \equiv t_{im} \pmod{d}$. From this it is easily seen that $t_{ij} = 0$ is an equivalence relation. Let us call such an equivalence class a *cyclic class*.

Since $t_{ij}+t_{jm} \equiv t_{im} \pmod{d}$, we see that $t_{ij} = t_{im}$ if and only if $t_{jm} = 0$, hence if and only if members j and m are in the same cyclic class. Let n be any integer. If $n \equiv t_{ij} \pmod{d}$, then the message originating from member i can only be in this one cyclic class after n steps. From

this it immediately follows that there are exactly d cyclic classes, and that the message moves cyclically from class to class, with cycle of length d. It is also easily seen that after sufficient time has elapsed, it can be in the hands of any member of the one cyclic class appropriate for n.

While this description of an equivalence class of the communication network holds in complete generality, the cycle degenerates when $d=1$. In this case there is a single "cyclic class," and after sufficient time has elapsed the message can be in the hands of any member at any time.

In particular, it is worth noting that if any member of the equivalence class can contact himself directly, then $d=1$. This is immediately seen from the fact that d is a divisor of any time in which a member can contact himself, and here d has to divide 1.

The number-theoretic result, § 1.4.1, is of such interest that its proof will be given here.

First of all we note that if the greatest common divisor d of the set is not 1, then we can divide all elements by d, and reduce the problem to the case $d=1$. Hence it suffices to treat this case. Here we have a set of numbers whose greatest common divisor is 1, and we must have a finite subset with this property. Hence, by a well-known result, there is a linear combination. $a_1n_1+a_2n_2+\cdots+a_kn_k$ of the elements (with positive or negative integers a_i) which is equal to 1. If we collect all the positive and all the negative terms separately, and remember that the set is closed under addition, we note that there must be elements m and n in the set, such that $m-n=1$ (m being the sum of the positive terms, and $-n$ the sum of the negative terms). Let q be any sufficiently large number, or more precisely $q \geqslant n(n-1)$. We can write $q=an+b$, where $a \geqslant (n-1)$ and $o \leqslant b \leqslant (n-1)$. Then we see that $q=(a-b)n+bm$, and hence q must be in the set.

§ 1.5 Probability measures. In making a probability analysis of an experiment there are two basic steps. First, a set of logical possibilities is chosen. This problem was discussed in § 1.2. Second a probability measure is assigned. The way that this second step is carried out will be discussed in this section. We consider first a finite possibility space. (For a more detailed discussion see FM Chapter IV or FMS Chapter III.)

1.5.1 DEFINITION. *Let* $\mathbf{U}=\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r\}$ *be a set of logical possibilities. A* probability measure for $\mathbf{U}$ *is obtained by assigning to each element* $\mathbf{a}_j$ *a positive number* $\mathbf{w}(\mathbf{a}_j)$, *called a weight, in such a way that the weights assigned have sum* 1. *The* measure *of a subset* $\mathbf{A}$ *of* $\mathbf{U}$, *denoted by* $\mathbf{m}(A)$, *is the sum of the weights assigned to elements of* $\mathbf{A}$.

1.5.2 THEOREM. *A probability measure* $\mathbf{m}$ *assigned to a possibility set* $\mathbf{U}$ *has the following properties:*

(1) For any subset $\mathbf{P}$ of $\mathbf{U}$, $0 \leqslant \mathbf{m}(\mathbf{P}) \leqslant 1$.

(2) If $\mathbf{P}$ and $\mathbf{Q}$ are disjoint subsets of $\mathbf{U}$, then $\mathbf{m}(\mathbf{P} \cup \mathbf{Q}) = \mathbf{m}(\mathbf{P}) + \mathbf{m}(\mathbf{Q})$.

(3) For any subsets $\mathbf{P}$ and $\mathbf{Q}$ of $\mathbf{U}$, $\mathbf{m}(\mathbf{P} \cup \mathbf{Q}) = \mathbf{m}(\mathbf{P}) + \mathbf{m}(\mathbf{Q}) - \mathbf{m}(\mathbf{P} \cap \mathbf{Q})$.

(4) For any set $\mathbf{P}$ in $\mathbf{U}$, $\mathbf{m}(\tilde{\mathbf{P}}) = 1 - \mathbf{m}(\mathbf{P})$.

1.5.3 DEFINITION. *Let* $\mathbf{p}$ *be a statement relative to a set* $\mathbf{U}$ *having truth set* $\mathbf{P}$. *The* probability of $\mathbf{p}$ relative to the probability measure $\mathbf{m}$ *is defined as* $\mathbf{m}(\mathbf{P})$.

In any discussion where there is a fixed probability measure we shall refer simply to the probability of $\mathbf{p}$ without mentioning each time the measure. From Theorem 1.5.2 and the relation of the connectives to the set operations, we have the following theorem:

1.5.4 THEOREM. *Let* $\mathbf{U}$ *be a set of possibilities for which a probability measure has been assigned. The probabilities of statements determined by this measure have the following properties:*

(1) *For any statement* $\mathbf{p}$, $0 \leqslant \mathbf{Pr}[\mathbf{p}] \leqslant 1$.

(2) *If* $\mathbf{p}$ *and* $\mathbf{q}$ *are inconsistent then* $\mathbf{Pr}[\mathbf{p} \vee \mathbf{q}] = \mathbf{Pr}[\mathbf{p}] + \mathbf{Pr}[\mathbf{q}]$.

(3) *For any two statements* $\mathbf{p}$ *and* $\mathbf{q}$, $\mathbf{Pr}[\mathbf{p} \vee \mathbf{q}] = \mathbf{Pr}[\mathbf{p}] + \mathbf{Pr}[\mathbf{q}] - \mathbf{Pr}[\mathbf{p} \wedge \mathbf{q}]$.

(4) *For any statement* $\mathbf{p}$, $\mathbf{Pr}[\sim \mathbf{p}] = 1 - \mathbf{Pr}[\mathbf{p}]$.

1.5.5 EXAMPLE. Given any finite set having s elements we can determine a probability measure by assigning weight $1/s$ to each element of $\mathbf{U}$. This measure is called the *equiprobable* measure. For any set $\mathbf{A}$ with r elements, $\mathbf{m}(\mathbf{A}) = r/s$. For example, this is the measure which would normally be assigned to the outcomes for the roll of a die. In this case $\mathbf{U} = \{1, 2, 3, 4, 5, 6\}$ and a weight of $1/6$ is assigned to each.

1.5.6 EXAMPLE. As an example of a situation where different weights would be assigned consider the following: A man observes a race between three horses $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. He feels that $\mathbf{a}$ and $\mathbf{b}$ have the same chance of winning but that $\mathbf{c}$ is twice as likely to win as $\mathbf{a}$. We take the possibility set to be $\mathbf{U} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and assign weights $\mathbf{w}(\mathbf{a}) = 1/4$, $\mathbf{w}(\mathbf{b}) = 1/4$ and $\mathbf{w}(\mathbf{c}) = 1/2$.

It is occasionally necessary to extend the above concepts to include the case of an experiment with an infinite sequence of possible outcomes. For example, consider the experiment of tossing a coin until the first

time that a head turns up. The possible outcomes would be $\mathbf{U}=\{1, 2, 3, \ldots\}$. The above definitions and theorems apply equally well to this possibility set. We will have an infinite number of weights assigned but we still must require that they have sum 1. In the example just mentioned we would assign weights $(1/2, 1/4, 1/8, \ldots)$. These weights form a geometric progression having sum 1.

§ 1.6 Conditional probability. It often happens that a probability measure has been assigned to a set **U** and then we learn that a certain statement **q** relative to **U** is true. With this new information we change the possibility set to the truth set **Q** of **q**. We wish to determine a probability measure on this new set from our original measure **m**. We do this by requiring that elements of **Q** should have the same relative weights as they had under the original assignment of weights. This means that our new weights must be the old weights multiplied by a constant to give them sum 1. This constant will be the reciprocal of the sum of the weights of all elements in **Q**, i.e. $1/\mathbf{m}(\mathbf{Q})$. (See FM Chapter IV or FMS Chapter III.)

1.6.1 DEFINITION. *Let* $\mathbf{U}=\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r\}$ *be a possibility set for which a measure has been assigned, determined by weights* $\mathbf{w}(\mathbf{a}_j)$. *Let* **q** *be a statement relative to* **U** (*not a self-contradiction*). *The* conditional probability measure given **q** *is a probability measure defined on* **Q** *the truth set of* **q**, *determined by weights*

$$\bar{\mathbf{w}}(\mathbf{a}_j) = \frac{\mathbf{w}(\mathbf{a}_j)}{\mathbf{m}(\mathbf{Q})}.$$

1.6.2 DEFINITION. *Let* **p** *and* **q** *be two statements relative to a set* **U** (**q** *not a self-contradiction*). *The conditional probability of* **p** *given* **q**, *denoted by* $\mathbf{Pr}[\mathbf{p}|\mathbf{q}]$ *is the probability of* **p** *computed from the conditional probability measure given* **q**.

1.6.3 THEOREM. *Let* **p** *and* **q** *be two statements relative to* **U** (**q** *not a self-contradiction*). *Assume that a probability measure* **m** *has been assigned to* **U**. *Then*

$$\mathbf{Pr}[\mathbf{p}|\mathbf{q}] = \frac{\mathbf{Pr}[\mathbf{p} \wedge \mathbf{q}]}{\mathbf{Pr}[\mathbf{q}]}$$

where $\mathbf{Pr}[\mathbf{p} \wedge \mathbf{q}]$ *and* $\mathbf{Pr}[\mathbf{q}]$ *are found from the measure* **m**.

1.6.4 EXAMPLE. In Example 1.5.6 assume that the man learns that horse **b** is not going to run. This causes him to consider the new possibility space $\mathbf{Q}=\{\mathbf{a}, \mathbf{c}\}$. The new weights which determine the conditional measure are $\bar{\mathbf{w}}(\mathbf{a})=\frac{1/4}{1/4+1/2}=1/3$ and $\bar{\mathbf{w}}(\mathbf{c})=\frac{1/2}{1/4+1/2}=2/3$.

We observe that it is still twice as likely that **c** will win than it is that **a** will win.

1.6.5 DEFINITION. *Two statements* **p** *and* **q** (*neither of which is a self-contradiction*) *are independent if* $\mathbf{Pr}[\mathbf{p} \wedge \mathbf{q}] = \mathbf{Pr}[\mathbf{p}] \cdot \mathbf{Pr}[\mathbf{q}]$.

It follows from Theorem 1.6.3 that **p** and **q** are independent if and only if $\mathbf{Pr}[\mathbf{p}|\mathbf{q}] = \mathbf{Pr}[\mathbf{p}]$ and $\mathbf{Pr}[\mathbf{q}|\mathbf{p}] = \mathbf{Pr}[\mathbf{q}]$. Thus to say that **p** and **q** are independent is to say that the knowledge that one is true does not effect the probability assigned to the other.

1.6.6 EXAMPLE. Consider two tosses of a coin. We describe the outcomes by $\mathbf{U} = \{\mathbf{HH}, \mathbf{HT}, \mathbf{TH}, \mathbf{TT}\}$. We assign the equiprobable measure. Let **p** be the statement "a head turns up on the first toss" and **q** the statement "a head turns up on the second toss." Then $\mathbf{Pr}[\mathbf{p} \wedge \mathbf{q}] = 1/4$, $\mathbf{Pr}[\mathbf{p}] = \mathbf{Pr}[\mathbf{q}] = 1/2$. Thus **p** and **q** are independent.

§ 1.7 Functions on a possibility space. Let $\mathbf{U} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r\}$ be a possibility space. Let **f** be a function with domain **U** and range $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_s\}$. That is, **f** assigns to each element **U** a unique element of **R**. If **f** assigns $\mathbf{r}_k$ to $\mathbf{a}_j$, we write $\mathbf{f}(\mathbf{a}_j) = \mathbf{r}_k$. We write $\mathbf{f} = \mathbf{r}_k$ for the statement "the value of the function is $\mathbf{r}_k$." This is a statement relative to **U**, since its truth value is known when the outcome $\mathbf{a}_j$ is known. Hence it has a truth set which is a subset of **U**. (See FMS Chapters II, III, or M⁴ Vol. II, Unit 1.)

1.7.1 DEFINITION. *Let* **f** *be a function with domain* **U** *and range* **R**. *Assume that a measure has been assigned to* **U**. *For each* $\mathbf{r}_k$ *in* **R** *let* $\mathbf{w}(\mathbf{r}_k) = \mathbf{Pr}[\mathbf{f} = \mathbf{r}_k]$. *The weights* $\mathbf{w}(\mathbf{r}_k)$ *determine a probability measure on the set* **R**, *called the* induced measure for **f**. *The weights are called the* induced weights.

We shall normally indicate the induced measure by giving both the range values and the weights in the form:

$$\mathbf{f}: \begin{Bmatrix} \mathbf{r}_1, & \mathbf{r}_2, & \ldots, & \mathbf{r}_s \\ \mathbf{w}(\mathbf{r}_1), & \mathbf{w}(\mathbf{r}_2), & \ldots, & \mathbf{w}(\mathbf{r}_s) \end{Bmatrix}.$$

Thus the induced weight of $\mathbf{r}_k$ in **R** is the measure of the truth set of $\mathbf{f} = \mathbf{r}_k$ in **U**.

1.7.2 EXAMPLE. In Example 1.6.6 let **f** be the function which gives the number of heads which turn up. The range of **f** is $\mathbf{R} = \{0, 1, 2\}$. The $\mathbf{Pr}[\mathbf{f} = 0] = 1/4$, $\mathbf{Pr}[\mathbf{f} = 1] = 1/2$, and $\mathbf{Pr}[\mathbf{f} = 2] = 1/4$. Hence the range and induced measure is:

$$\mathbf{f}: \begin{Bmatrix} 0 & 1 & 2 \\ 1/4 & 1/2 & 1/4 \end{Bmatrix}.$$

1.7.3 DEFINITION. *Let* **U** *be a possibility space, and* **f** *and* **g** *be two functions with domain* **U**, *each having as range a set of numbers. The function* $\mathbf{f}+\mathbf{g}$ *is the function with domain* **U** *which assigns to* $\mathbf{a}_j$ *the number* $\mathbf{f}(\mathbf{a}_j)+\mathbf{g}(\mathbf{a}_j)$. *The function* $\mathbf{f}\cdot\mathbf{g}$ *is the function with domain* **U** *which assigns to* $\mathbf{a}_j$ *the number* $\mathbf{f}(\mathbf{a}_j)\cdot\mathbf{g}(\mathbf{a}_j)$. *For any number* c *the* constant function **c** *is the function which assigns the number* c *to every element of* **U**.

Let **U** be a possibility space for which a measure has been assigned. Then if **f** and **g** are two numerical functions with domain **U**, $\mathbf{f}+\mathbf{g}$ and $\mathbf{f}\cdot\mathbf{g}$ will be functions with domain **U**, and as such have induced measures. In general there is no simple connection between the induced measures of these functions and the induced measure for **f** and **g**.

1.7.4 EXAMPLE. In Example 1.6.6 let **g** be a function having the value 1 if a head turns up on the first toss and 0 otherwise. Let **h** be a function having the value 1 if a head turns up on the second toss and 0 if a tail turns up. Then the range and induced measures for **g**, **h**, $\mathbf{g}+\mathbf{h}$, and $\mathbf{g}\cdot\mathbf{h}$ are

$$\mathbf{g}: \begin{Bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{Bmatrix}$$

$$\mathbf{h}: \begin{Bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{Bmatrix}$$

$$\mathbf{g}+\mathbf{h}: \begin{Bmatrix} 0 & 1 & 2 \\ 1/4 & 1/2 & 1/4 \end{Bmatrix}$$

$$\mathbf{g}\cdot\mathbf{h}: \begin{Bmatrix} 0 & 1 \\ 3/4 & 1/4 \end{Bmatrix}.$$

1.7.5 DEFINITION. *Let* **f** *be a function defined on* **U**. *Let* **p** *be a statement relative to* **U** *having truth set* **P**. *Assume that a measure* **m** *has been assigned to* **U**. *Let* $\mathbf{f}'$ *be the function* **f** *considered only on the set* **P**. *Then the induced measure for* $\mathbf{f}'$ *calculated from the conditional measure given* **p** *is called the* conditional induced measure for **f** given **p**.

1.7.6 DEFINITION. *Let* **f** *and* **g** *be two functions defined on a space* **U** *for which a probability measure has been assigned. Then* **f** *and* **g** *are* independent *if, for any* $\mathbf{r}_k$ *in the range of* **f** *and* $\mathbf{s}_j$ *in the range of* **g**, *the statements* $\mathbf{f}=\mathbf{r}_k$ *and* $\mathbf{g}=\mathbf{s}_j$ *are independent statements.*

An equivalent way to state the condition for independence of two