CONDITIONAL INDEPENDENCE IN APPLIED PROBABILITY

Paul E. Pfeiffer
Department of Mathematical Sciences
Rice University
Houston, Texas

edc/umap/55 chapel st./newton, mass. 02160

umap
Modules and Monographs in
Undergraduate Mathematics
and its Applications Project

CONDITIONAL INDEPENDENCE IN APPLIED PROBABILITY

Paul E. Pfeiffer
Department of Mathematical Sciences
Rice University
Houston, Texas

The Project acknowledges Robert M. Thrall, Chairman of the UMAP Monograph Editorial Board, for his help in the development and review of this monograph.

Modules and Monographs in Undergraduate Mathematics and its Applications Project

The goal of UMAP is to develop, through a community of users and developers, a system of instructional modules and monographs in undergraduate mathematics and its applications which may be used to supplement existing courses and from which complete courses may eventually be built.

The Project is guided by a National Steering Committee of mathematicians, scientists, and educators. UMAP is funded by a grant from the National Science Foundation to Education Development Center, Inc., a publicly supported, nonprofit corporation engaged in educational research in the U.S. and abroad.

UMAP wishes to thank Charles Harvey of Rice University for his review of this manuscript.

The Project acknowledges the help of the Monograph Editorial Board in the development and review of this monograph. Members of the Monograph Editorial Board include:

ISBN-13: 978-1-4612-6337-1 e-ISBN-13: 978-1-4612-6335-7
DOI: 10.1007/978-1-4612-6335-7

CONTENTS

Preface

It would be difficult to overestimate the importance of stochastic independence in both the theoretical development and the practical applications of mathematical probability. The concept is grounded in the idea that one event does not "condition" another, in the sense that occurrence of one does not affect the likelihood of the occurrence of the other. This leads to a formulation of the independence condition in terms of a simple "product rule," which is amazingly successful in capturing the essential ideas of independence.

However, there are many patterns of "conditioning" encountered in practice which give rise to quasi independence conditions. Explicit and precise incorporation of these into the theory is needed in order to make the most effective use of probability as a model for behavioral and physical systems. We examine two concepts of conditional independence.

The first concept is quite simple, utilizing very elementary aspects of probability theory. Only algebraic operations are required to obtain quite important and useful new results, and to clear up many ambiguities and obscurities in the literature.

The second concept of conditional independence has been employed for some time in advanced treatments of Markov processes. Couched in terms of the abstract notion of conditional expectation, given a sigma field of events, this concept has been available only to those with the requisite measure-theoretic preparation. Since the use of this concept in the theory of Markov processes not only yields important mathematical results, but also provides conceptual advantages for the modeler, it should be made available to a wider class of users. The case is made more compelling

by the fact that the concept, once available, has served to provide new precision and insight into the handling of a number of topics in probable inference and decision, not related directly to Markov processes.

The reader is assumed to have the background provided by a good undergraduate course in applied probability (see Secs A1, A2). Introductory courses in calculus, linear algebra, and perhaps some differential equations should provide the requisite experience and proficiency with mathematical concepts, notation, and argument. In general, the mathematical maturity of a junior or senior student in mathematical sciences, engineering, or one of the physical sciences should be adequate, although the reader need not be a major in any of these fields.

Considerable attention is given to careful mathematical development. This serves two types of interests, which may enhance and complement one another. The serious practitioner of the art of utilizing mathematics needs insight into the system he is studying. He also needs insight into the model he is using. He needs to distinguish between properties of the model which are definitive or axiomatic (and hence appear as basic assumptions) and those which are logical consequences (i.e., theorems) deduced from the axiomatic properties. For example, if his experience makes it reasonable to assume that a dynamic system is characterized by lack of "memory", so that the future is conditioned only by the present state and not past history, then it is appropriate to consider representing the system as a Markov process. Should the system fail to exhibit certain consequences of the Markov assumption, then that fundamental assumption must be reexamined. The distinction between fundamental properties and derived properties is an aid to efficient and intelligent use of mathematics (as well as insurance against contradictory assumptions).

The serious mathematician who wishes to enlarge his knowledge and appreciation of the applications of mathematics (and perhaps discover new, significant problems) may be deterred by the inadequate articulation of mathematics in much of the applied literature. This may be a serious barrier to what should be a cooperative endeavor. Hopefully, the present treatment will help remove any such barrier to consideration of the interesting and important topic of conditional independence.

In order to recast the theory of conditional independence of random vectors in more elementary terms, it has been necessary to extend the usual introductory treatment of conditional expectation, given a random vector. The treatment intends to bridge the gap between the usual intuitive introductory treatment, based on a concept of conditional distribution, and a more general approach found in advanced, measure-theoretic treatments. Because of the importance of conditional expectation as a tool in the study of random processes and of decision theory, the results should be useful beyond the scope of the present investigation.

Acknowledgements

It is apparent that a work of this sort draws on a variety of sources, many of which are no longer identifiable. Much of the impetus for writing came from teaching courses in probability, random processes, and operations research. The response of students and colleagues to various presentations has been helpful in many ways. The development of the concept of conditional independence of events has been stimulated and shaped in large part by my collaboration with David A. Schum, Professor of Psychology, in some aspects of his work on human inference. He has read critically several versions of the manuscript. Charles M. Harvey of Dickinson College, while on visiting appointment in Mathematical Sciences at Rice University, read

critically a preliminary manuscript presented for review. His comments were helpful in planning the final, extensively revised manuscript. Dr. David W. Scott of Baylor College of Medicine and Rice University used some of the results in recent work. His comments were helpful in improving exposition at several points, and his work provided an interesting applications problem.

Paul E. Pfeiffer

A. Preliminaries

A. PRELIMINARIES

CONDITIONAL INDEPENDENCE IN APPLIED PROBABILITY

A. Preliminaries

In this monograph, we assume the reader has reasonable facility with elementary probability at the level of such texts as Pfeiffer and Schum [1973], Ash [1970], or Chung [1974]. In particular, we suppose the reader is familiar with the concept of a random variable, or a random vector, as a mapping from the basic space to the real line R, or to Euclidean space R^n, and with the notion of mathematical expectation and its basic properties (cf Pfeiffer and Schum [1973], Chaps 8, 10, 13). In the following sections, we summarize various fundamental concepts and results in a form, terminology, and notation to be utilized in subsequent developments. In some cases, we simply express familiar material in a form useful for our purposes; in others, we supplement the usual introductory treatment, especially with an informal presentation of certain ideas and results from measure theory. The reader may wish to scan this material rapidly, returning as needed for later reference.

1. Probability spaces and random vectors

A probability space, or probability system, consists of a triple $(\Omega,\mathfrak{F},P)$.

1) Ω is the basic space, or sample space, each element of which represents one of the conceptually possible outcomes of a specified trial, or experiment. Each elementary outcome ω is an element of the basic space Ω.

2) $\mathfrak{F}$ is a class of subsets of Ω. Each of the subsets in this class is an event. The event A occurs iff the ω resulting from the trial is an element of A. Since it is desirable that the sets formed by complements, countable unions, or countable intersections of events also be events, the class $\mathfrak{F}$ must have the properties of a sigma

field (also called a Borel field or a sigma algebra) of sets.

3) The probability measure P assigns to each event A a number $P(A)$ in such a manner that three basic axioms (and logical consequences) hold: i) $P(A) \geq 0$, ii) $P(\Omega) = 1$, and iii) P is countably additive.

We utilize standard notation for the empty set (impossible event), complements, unions, and intersections. Thus, for example,

$\emptyset$ is the empty set (the impossible event),

$\bigcup_{i=1}^{\infty} A_i$ is the union of the infinite class $\{A_i : 1 \leq i < \infty\}$,

$\bigcap_{i=1}^{n} B_i$ is the intersection of the finite class $\{B_i : 1 \leq i \leq n\}$,

A^c is the complement of the set A.

In addition, we employ the notation $\biguplus_{i=1}^{n} B_i$ to indicate not only that we have taken the union of the class $\{B_i : 1 \leq i \leq n\}$, but also that the class is disjoint (the events are mutually exclusive). Thus, the expression

$A = \biguplus_{i=1}^{\infty} A_i$ means the same as the pair of statements

i) $A = \bigcup_{i=1}^{\infty} A_i$ and ii) $A_i A_j = \emptyset$ for $i \neq j$.

A random vector is viewed as a mapping from the basic space Ω to n-dimensional Euclidean space $\mathbb{R}^n$. For $n = 1$, we have a real-valued random variable. A random vector $X: \Omega \to \mathbb{R}^n$ may be considered to be the joint mapping $(X_1, X_2, \ldots, X_n): \Omega \to \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$ produced by the coordinate random variables $X_1, X_2, \ldots, X_n$.

Since we want to be able to make probability statements about possible sets of values to be taken on by random vectors, we must introduce measurability considerations. In the real-valued case ($n = 1$), we should like to speak of the probability that X takes on a value no greater than some real number t. Since probability is assigned to events, the

set $\{\omega: X(\omega) \leq t\}$ should be an event for any real number t. This may be viewed schematically with the aid of a mapping diagram, as in Figure A1-1. We are interested in the set A of those elementary outcomes ω which are mapped into the interval $I_t = (-\infty, t]$. Since we also want to consider complements, countable unions, and countable intersections of such events, we must consider complements, countable unions, and countable intersections of such intervals on the real line. We are thus led to consider the minimal sigma field $\mathcal{B}$ of subsets of the real line which includes all the semi-infinite intervals of the form $I_t = (-\infty, t]$. This is the class $\mathcal{B}$ of Borel sets on the real line. A similar consideration leads to defining the class $\mathcal{B}$ of Borel sets on R^n as the minimal sigma field which includes all semi-infinite intervals of the form $I(t_1, t_2, \ldots, t_n) = (-\infty, t_1] \times (-\infty, t_2] \times \ldots \times (-\infty, t_n]$. We say that $X: \Omega \to R^n$ is a random vector iff $X^{-1}(M) = \{\omega: X(\omega) \in M\}$ is an event for each Borel set M in R^n. A standard result of measure theory, which we assume without proof, is that $X^{-1}(M)$ is an event for each Borel set M iff $X^{-1}[I(t_1, t_2, \ldots, t_n)]$ is an event for each n-tuple $(t_1, t_2, \ldots, t_n)$ of real numbers (i.e., for each element of R^n). Real-valued random variables are included as the special case $n = 1$.

It is an easy consequence of elementary mapping theorems that the class $\mathcal{F}(X)$ of all inverse images $X^{-1}(M)$ of Borel sets is a sigma field. We refer to this class as the sigma field determined by X. It must be a subclass of the class $\mathcal{F}$ of events in order for X to be a random vector.

We often need to consider functions of random vectors. If $X: \Omega \to R^n$ and $g: R^n \to R^m$, then $Z = g \circ X = g(X)$ is a function $\Omega \to R^m$. If g has the property that $N = g^{-1}(M)$ is a Borel set in R^n for each Borel set M in its codomain R^m, then Z is a random vector, since $Z^{-1}(M) =$

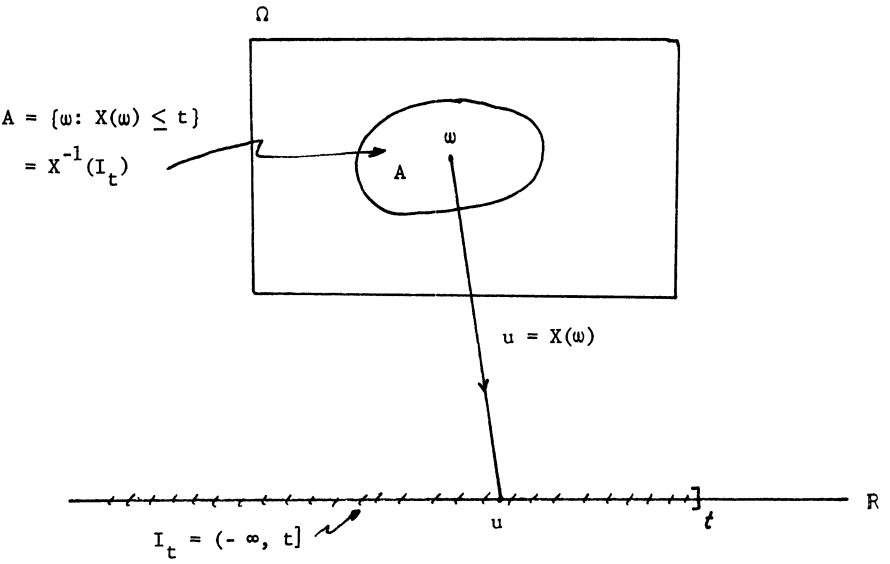


Figure A1-1. Mapping diagram with inverse image of semi-infinite interval I_t.

$X^{-1}g^{-1}(M) = X^{-1}(N)$ is an event. Thus, each event determined by Z is an event determined by X. This may be expressed by the relation $\mathcal{F}(Z)$ is contained in $\mathcal{F}(X)$. This condition is oftened indicated by saying that Z is measurable with respect to X (or Z is measurable-X). A function g with the mapping property described above is known as a Borel function. From somewhat advanced arguments, it is known that if Z is measurable-X, then there is a Borel function g such that $Z = g \circ X = g(X)$. We assume this important result without proof.

We have introduced the class of Borel functions in a somewhat abstract manner to solve the problem of when a function of a random vector is itself a random vector. But how do we know whether or not a function encountered in practice is Borel? It turns out that almost any function $g: R^n \to R^m$ which we may want to consider is Borel. For this reason, in many introductory treatment little or nothing is said about Borel functions.

Borel functions constitute a generalization of the class of continuous functions. Continuous functions have the property that the inverse image of any open set is open. It is known that the class of Borel sets on R^n is the minimal sigma field which includes all open sets in R^n. From this fact it may be shown that any continuous function from R^n to R^m is Borel. Any piecewise continuous real function $g: R \to R$ is Borel. Linear combinations, products, and compositions (functions of functions) of Borel functions are Borel. If $\{g_n: 1 \leq n\}$ is a sequence of Borel functions from R^n to R^m which converge for each t in R^n, the limit function g is a Borel function.

The indicator function I_A for set A in Ω, defined by $I_A(\omega) = 1$ for ω in A and zero otherwise, is particularly useful. If A is an event, I_A is a random variable. Indicator functions may be defined, as

well, on R^n. If M is a Borel set in R^n, then I_M is a Borel function from R^n to R. If c is an element of R^m, then cI_M is a Borel function from R^n to R^m. If X is a random vector and M is a Borel set on the codomain of X, then $I_M(X)$ is a real-valued random variable, measurable-X. If M is a subset of R^m and N is a subset of R^n, then the cartesian product $M \times N = \{(t,u): t \in M, u \in N\}$ is a subset of $R^m \times R^n$. The indicator function $I_{M \times N}: R^m \times R^n \to R$ satisfies the equation

$$I_{M \times N}(t,u) = I_M(t)I_N(u) \quad \forall\ t \in R^m,\ u \in R^n$$

since $(t,u) \in M \times N$ iff both $t \in M$ and $u \in N$.

The following result is basic in the development of the concept of conditional expectation.

Theorem A1-1

a) If Y is a random vector with codomain R^m, M is any Borel set in R^m, and $C = \{\omega: Y(\omega) \in M\} = Y^{-1}(M)$, then $I_C = I_M(Y)$.

b) If g is a Borel function $R^m \to R^n$ and $Z = g(Y)$, then for any Borel set N in R^n, there is a Borel set $M = g^{-1}(N)$ in R^m such that $I_N(Z) = I_M(Y)$.

PROOF

a) $I_M[Y(\omega)] = 1$ iff $Y(\omega) \in M$ iff $\omega \in C$ iff $I_C(\omega) = 1$

b) The relation $C = Y^{-1}(M) = Z^{-1}(N)$ is an elemental property of composite mappings. By a), $I_N(Z) = I_C = I_M(Y)$ []

The indicator function is useful in representing discrete random variables, which take on a finite or countably infinite set of values. In the finite case, the term simple random variable is commonly used. Suppose the range (set of possible values) of X is $S = \{t_1, t_2, \ldots, t_N\} \subset R^m$. Let $A_i = \{\omega: X(\omega) = t_i\}$. Then the class $\{A_i: 1 \leq i \leq N\}$ is a partition,

and $X = \sum_{i=1}^{N} t_i I_{A_i}$. We refer to this representation as canonical form (if one of the values is zero, we include a term with zero coefficient). It is easy to show that any real random variable is the limit of a sequence of such simple random variables (the sequence is not unique). If X is nonnegative, it is the limit of an increasing sequence of nonnegative, simple random variables (cf Pfeiffer and Schum [1973], Sec 8.8). Similar statements may be made about Borel functions. A simple Borel function $g: \mathbb{R}^m \to \mathbb{R}^n$ has canonical form $g = \sum_{i=1}^{N} t_i I_{M_i}$, where $t_i \in \mathbb{R}^n$ and each $M_i = \{u \in \mathbb{R}^m: g(u) = t_i\}$ is a Borel set in $\mathbb{R}^m$.

A random vector induces a probability distribution on the Borel sets of its codomain. To each Borel set M is assigned the probability mass on the event $X^{-1}(M)$. A probability measure P_X is defined on the Borel sets by the assignment $P_X(M) = P[X^{-1}(M)] = P(X \in M)$. This is a true probability measure, with the Borel sets serving as events. This mass distribution may also be described by a probability distribution function F_X or, in suitable cases, by a probability density function f_X. These matters are assumed to be familiar.

For many purposes, if a random vector is modified on a set of ω having zero probability, no significant difference is realized in probability calculations. For example, if X and Y are two real random variables with the property that the set of ω for which $X(\omega) \neq Y(\omega)$ has probability zero, these random variables have the same mathematical expectation.

DEFINITION. Random vectors X, Y are almost surely equal, denoted $X = Y$ a.s., iff they have the same codomain and the set $\{\omega: X(\omega) \neq Y(\omega)\}$ has probability zero.

More generally, a relation between random vectors is said to hold almost surely (a.s.), or to hold for almost every (a.e.) ω, iff the set of

ω for which the relation fails to hold has probability zero.

We are frequently concerned with functions of random vectors. Suppose we have random vector $X: \Omega \to R^n$ and have two Borel functions $g, h: R^n \to R^m$. If these functions have the property that $g(t) = h(t)$ for all t on the range of X, then we must have $g[X(\omega)] = h[X(\omega)]$ for all ω. Again, we may not need this equality for all ω. It may be sufficient to have equality for almost every ω (i.e., for all ω except possibly an exceptional set of probability zero). Suppose $M_0 = \{t \in R^n: g(t) \neq h(t)\}$. Then $g[X(\omega)] \neq h[X(\omega)]$ iff $X(\omega)$ is one of the values in M_0. Hence, $g(X) = h(X)$ a.s. iff the set of ω for which $X(\omega) \in M_0$ has probability zero. But this is just the condition that the induced probability $P_X(M_0) = P(X \in M_0) = 0$.

The notion of almost-sure equality for random vectors can be extended to Borel functions when the probability measure is defined on the class of Borel sets on the domain of the functions. We are particularly interested in the case that such measures are probability measures induced by random vectors.

DEFINITION. If g, h are Borel functions from R^n to R^m and P_X is a probability measure on the Borel sets on R^n, then g and h are said to be <u>almost surely equal</u> $[P_X]$ iff the set $M_0 = \{t \in R^n: g(t) \neq h(t)\}$ satisfies the condition $P_X(M_0) = 0$.

The discussion above provides the justification for the following

<u>Theorem</u> A1-2

$g(X) = h(X)$ a.s. iff $g = h$ a.s. $[P_X]$, where P_X is the probability measure induced by the random vector X. []

Independence of random vectors is expressed in terms of the events they determine.

DEFINITION. An arbitrary class $\{X_i : i \in J\}$ of random vectors is <u>independent</u> iff for each class $\{M_i : i \in J\}$ of Borel sets on the respective codomains of the X_i the class $\{X^{-1}(M_i) : i \in J\}$ of events is independent.

This means that the product rule holds for each finite subclass of the class of events. The following is known to be consistent with the above.

DEFINITION. Two classes $\{X_t : t \in T\}$ and $\{Y_u : u \in U\}$ form an <u>independent family</u> of classes iff for each finite $T_n \subset T$ and $U_m \subset U$ the random vectors $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and $(Y_{u_1}, Y_{u_2}, \ldots, Y_{u_m})$ form an independent pair.

The latter definition extends readily to arbitrary families of classes. In the next section, we state the condition for independence of a class of random vectors in terms of mathematical expectation.

If $\{X,Y\}$ is an independent pair of random vectors (any finite dimensions) and g, h are Borel functions on the codomains of X, Y, respectively, then $\{g(X),h(Y)\}$ is an independent pair. This follows from the fact that $\{g(X) \in M\} = \{X \in g^{-1}(M)\}$ and $\{h(Y) \in N\} = \{Y \in h^{-1}(N)\}$, so that $P(\{g(X) \in M\} \cap \{h(Y) \in N\} = P(\{X \in g^{-1}(M)\} \cap \{Y \in h^{-1}(N)\}) = P[X \in g^{-1}(M)]P[Y \in h^{-1}(N)] = P[g(X) \in M]P[h(Y) \in N]$. It should be apparent how this result extends to arbitrary classes.

2. Mathematical expectation

The concept of mathematical expectation incorporates the notion of a probability weighted average. Suppose X is a simple, real-valued random variable with range $\{t_1, t_2, \ldots, t_n\}$. The mathematical expectation of X is $E[X] = \sum_{i=1}^{n} t_i P(X = t_i)$. Each possible value t_i is weighted by the probability that value will be realized; these weighted values are summed to give a probability weighted sum; since the total weight is one, the sum is the same as the average.

To extend the notion, we consider next a nonnegative random variable X. In this case, there is a nondecreasing sequence of simple random variables which converge to X. We define

$$E[X] = \int X \, dP = \lim_n E[X_n].$$

A study of the technical details shows that the limit does not depend upon the particular approximating sequence selected. To complete the extension to the general case, we represent X as the difference $X_+ - X_-$ of the two nonnegative random variables defined as follows:

$$X_+(\omega) = \begin{cases} X(\omega) & \text{for } X(\omega) \geq 0 \\ 0 & \text{for } X(\omega) < 0 \end{cases} \qquad X_-(\omega) = \begin{cases} 0 & \text{for } X(\omega) \geq 0 \\ -X(\omega) & \text{for } X(\omega) < 0. \end{cases}$$

Then $E[X] = E[X_+] - E[X_-]$. Thus $E[X]$ is the limit of the probability weighted averages of the values of the approximating simple functions. As such, mathematical expectation should have properties of sums or averages which "survive passage to a limit." This is, in fact, the case. The defining procedure defines a very general type of integration (Lebesgue integration).

For convenience, we list and assign numbers to those properties of mathematical expectation which are most useful in investigations such as those in subsequent sections. Since an indicator function for an event

is a simple random variable whose range is $\{0, 1\}$, we have

E1) $E[I_A] = P(A)$.

Use of Theorem A1-1 and the fact that $I_{M \times N}(X,Y) = I_M(X) I_N(Y)$ gives the following important special cases.

E1a) $E[I_M(X)] = P(X \in M)$ and $E[I_M(X) I_N(Y)] = P(X \in M, Y \in N)$ (with extension by mathematical induction to any finite number of random vectors).

Elementary arguments show that the following properties of sums hold also for mathematical expectation in general.

E2) Linearity. $E[aX + bY] = aE[X] + bE[Y]$ (with extension by mathematical induction to any finite linear combination).

E3) Positivity; monotonicity.

a) $X \geq 0$ a.s. implies $E[X] \geq 0$, with equality iff $X = 0$ a.s.

b) $X \geq Y$ a.s. implies $E[X] \geq E[Y]$, with equality iff $X = Y$ a.s.

It should be noted that monotonicity follows from linearity and positivity.

The next property is not ordinarily discussed in elementary treatments. However, it is essential to much of the theory of mathematical expectation. Suppose $X_n \leq X_{n+1}$ a.s. for all $n \geq 1$ and $X_n(\omega) \to X(\omega)$ for a.e. ω. By property E3), we must have $E[X_n] \leq E[X_{n+1}] \leq E[X]$. Since a bounded monotone sequence of real numbers always converges, we must have $\lim_{n\to\infty} E[X_n] = L \leq E[X]$. Sophisticated use of elementary ideas establishes the fact that the limit $L = E[X]$. A similar argument holds for monotone decreasing sequences. Thus, we have

E4) Monotone convergence. If $X_n \to X$ monotonically a.s., then $E[X_n] \to E[X]$ monotonically.

In many ways, these four properties characterize mathematical expectation as an integral. A surprising number of other properties stem from these. In the development of the idea of conditional expectation, we establish its integral-like character by establishing analogs of E1) through E4).

By virtue of the definition and property E1a) we can characterize independence of random vectors as follows.

E5) Independence. The pair $\{X,Y\}$ of random vectors is independent

iff $E[I_M(X)I_N(Y)] = E[I_M(X)]E[I_N(Y)]$ for all Borel sets M, N on the codomains of X, Y, respectively,

iff $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for all real-valued Borel functions g, h such that the expectations exist.

For an arbitrary family of random vectors, we have independence iff such a product rule holds for every finite subclass of two or more members.

The next property plays an essential role in the development of the concept of conditional expectation. We prove the basic result, which suffices for developing the properties of conditional expectation; the extension, whose proof requires some advanced ideas from measure theory, is used in developing certain equivalent conditions for conditional independence, given a random vector (Sec D5).

E6) Uniqueness.

a) Suppose Y is a random vector with codomain R^m and g, h are real-valued Borel functions on the range of Y. If $E[I_M(Y)g(Y)] = E[I_M(Y)h(Y)]$ for all Borel sets M in the codomain of Y, then $g(Y) = h(Y)$ a.s.

b) More generally, if $E[I_M(Y)I_N(Z)g(Y,Z)] = E[I_M(Y)I_N(Z)h(Y,Z)]$ for all Borel sets M, N in the codomains of Y, Z, respectively, then $g(Y,Z) = h(Y,Z)$ a.s.

PROOF OF a).

Suppose $g(u) > h(u)$ for u in the set N. Then $I_N(Y)g(Y) \geq I_N(Y)h(Y)$, with equality iff $Y(\omega)$ does not belong to N. By E3), $E[I_N(Y)g(Y)] = E[I_N(Y)h(Y)]$ iff $I_N(Y)g(Y) = I_N(Y)h(Y)$ a.s. iff $P(Y \in N) = 0$.

A similar argument holds for the opposite inequality. Thus, the total probability of the event $\{g(Y) \neq h(Y)\}$ is zero.

DISCUSSION OF b)

The second part is more general, since the sets $Q = M \times N$, with $I_Q = I_M I_N$, form only a subclass of the Borel sets on the codomain of the combined vector (X,Y). However, a standard type of argument in measure theory shows that if equality holds for sets of this subclass, it must hold for all Borel sets. Application of part a) gives the desired result. []

Several useful properties are based on E1 through E4), with monotone convergence playing a key role. The following are among the most important.

E7) Fatou's lemma. If $X_n \geq 0$ a.s., $E[\liminf X_n] \leq \liminf E[X_n]$.

E8) Dominated convergence. If $X_n \to X$ a.s. and $|X_n| \leq Y$ a.s., for each n, with $E[Y]$ finite, then $E[X_n] \to E[X]$.

E9) Countable additivity. Suppose $E[X]$ exists and $A = \biguplus_{i=1}^{\infty} A_i$. Then $E[I_A X] = \sum_{i=1}^{\infty} E[I_{A_i} X]$.

The following property is used as the basis for a general definition of conditional expectation, given a random vector. It is based on the celebrated Radon-Nikodym theorem and the fact, noted in the previous section, that if Z is measurable-Y, then there is a Borel function e such that $Z = e(Y)$. We accept this result without proof. It is made plausible in certain special cases in the developments in Sec C2.

E10) Existence. If $E[g(X)]$ is finite, then there is a real-valued Borel function e, unique a.s. $[P_Y]$, such that $E[I_M(Y)g(X)] = E[I_M(Y)e(Y)]$ for all Borel sets M in the codomain of Y.

Recall, by Theorem A1-2, e is unique a.s. $[P_Y]$ iff $e(Y)$ is unique a.s.

A number of standard inequalities are employed repeatedly in probability theory. Establishment of these depends upon setting up the appropriate

inequalities on random variables, then utilizing monotonicity E3). The appropriate inequalities on the random variables are often expressions of classical inequalities in ordinary analysis. Some of the more important inequalities are listed for convenient reference in Appendix I.

3. Problems

A-1. For each of the following random variables, describe the sigma field $\mathcal{F}(X)$ determined by X.

i) $X = I_A$

ii) $X = aI_A + bI_B + cI_C$ (canonical form)

A-2. If $X = -2I_A + 0I_B + I_C + 4I_D$ (canonical form), describe $X^{-1}(M)$ for

i) $M = (-\infty, 0]$, ii) $M = (-2, 1] \uplus (2,4]$. iii) $M = (-\infty, 3]$

A-3. Suppose X has distribution function F_X with

$$F_X(t) = \begin{cases} 0 & \text{for } t < 0 \\ (1 + 3t)/4 & \text{for } 0 \leq t < 1 \\ 1 & \text{for } 1 \leq t \end{cases}$$

For which of the following functions, if any, is $g_i = g_k$ a.s. $[P_X]$?

$g_1(t) = t + 1$ for all t

$$g_2(t) = \begin{cases} 0 & \text{for } t \leq 0 \\ t + 1 & \text{for } 0 < t < 1 \\ 2 & \text{for } 1 \leq t \end{cases}$$

$g_3(t) = t + k + 1$ for $k \leq t < k + 1$, all integers k, all t

A-4. If X and Y are real random variables, let

$$X_+(\omega) = \begin{cases} X(\omega) & \text{for } X(\omega) \geq 0 \\ 0 & \text{for } X(\omega) < 0 \end{cases} \qquad X_-(\omega) = \begin{cases} -X(\omega) & \text{for } X(\omega) \leq 0 \\ 0 & \text{for } X(\omega) > 0 \end{cases}$$

Show that

a) X_+ and X_- are Borel functions of X, hence are random variables.

b) XY is a random variable

c) $aX + bY$ is a random variable (a,b are constants).

A-5. Suppose $g: R^m \to R^n$ and $f: R^n \to R^q$ are Borel functions. Show that the composition $f \circ g: R^m \to R^q$ is Borel.

A-6. Use Theorem A1-1 and property E1) for expectation to establish property E1a).

A-7. Use linearity E2) and positivity for expectation to establish monotonicity.

A-8. a) Suppose $X \geq 0$ and $E[X]$ is finite. Use the monotone convergence theorem E4) to establish countable additivity E9) for expectation.

b) Extend the result of part a) to the general case.

A-9. If X is real, use the fact that $X \leq |X|$ and $-X \leq |X|$ to establish the triangle inequality E11) for expectation.

A-10. Establish the mean-value theorem E12) for expectation.

B. Conditional Independence of Events

B. CONDITIONAL INDEPENDENCE OF EVENTS

B. Conditional independence of events

1. The concept

In setting up a probability model for a system under study, the modeler utilizes all available prior knowledge about the system to determine probability assignments to appropriate events. This knowledge may be obtained from systematic statistical study, or from mathematical deductions based on assumptions supported by experience or experiment, or, less formally, from the judgment of a decision maker. These probability assignments serve to determine a prior probability measure. The probability $P(A)$ of an event A provides a measure of the likelihood of the occurrence of this event.

Further experience or experiment may produce information which makes it appropriate to revise the probability assignments to reflect new likelihoods of various events. Such revisions amount to the introduction of a new probability measure. Typically, the information received yields partial knowledge of the character of the outcome. When properly expressed, this new information serves to identify an event C which has occurred. There may be subtleties and difficulties in determining exactly what this conditioning event C is (cf. Pfeiffer and Schum [1973], Sec 5-1). The difficulties center about the question: What information is obtained by whom? But, in principle at least, such an event is determined.

There is nothing in the probability model to require a specific manner of reassigning probabilities. However, considerable experience has shown that a fruitful way to make the new assignment of probability to event A, given the occurrence of conditioning event C, is to utilize the rule

$$P(A|C) = P(AC)/P(C), \quad \text{provided, of course,} \quad P(C) > 0.$$

We call $P(A|C)$ the conditional probability of A, given C. For fixed C, $P(\cdot|C)$ is a new probability measure, with all the formal properties

of the original, or prior, probability measure $P(\cdot)$.

It sometimes happens that occurrence of the event C does not affect the likelihood that A will (or will not) occur. Thus, we may be able to assert that $P(A|C) = P(A)$ or $P(A|C) = P(A|C^c)$. As a matter of fact, straightforward use of the defining relation for conditional probability shows that if $0 < P(A) < 1$ and $0 < P(C) < 1$, then the following sixteen relations are equivalent-- that is, if one holds, so do the others.

$$\begin{array}{lll} P(A|C) = P(A) & P(C|A) = P(C) & P(AC) = P(A)P(C) \\ P(A|C^c) = P(A) & P(C^c|A) = P(C^c) & P(AC^c) = P(A)P(C^c) \\ P(A^c|C) = P(A^c) & P(C|A^c) = P(C) & P(A^cC) = P(A^c)P(C) \\ P(A^c|C^c) = P(A^c) & P(C^c|A^c) = P(C^c) & P(A^cC^c) = P(A^c)P(C^c) \end{array}$$

$$P(A|C) = P(A|C^c) \quad P(A^c|C) = P(A^c|C^c) \quad P(C|A) = P(C|A^c) \quad P(C^c|A) = P(C^c|A^c).$$

If any of these holds, we suppose the events A, C form an independent pair, in a probabilistic sense. It is easy to check that the equivalence of the four product rules in the right-hand column holds for the cases in which either $P(A)$ or $P(C)$ takes one of the extreme values 0 or 1. Also, the first product rule is symmetric with respect to the events A, C. Thus, it is convenient to make the definition of independence in terms of this product rule, as follows:

DEFINITION. The pair $\{A,B\}$ of events is (<u>stochastically</u>) <u>independent</u> iff the product rule $P(AB) = P(A)P(B)$ holds.

An arbitrary class of events is independent iff a corresponding product rule holds for every finite subclass of two or more events from the class. The list of equivalent relations above (with C replaced by B) shows that

If any one of the pairs $\{A,B\}$, $\{A,B^c\}$, $\{A^c,B\}$, or $\{A^c,B^c\}$ is independent, so are the others.

Although the product rule is the basis of the formal definition, the essential idea of independence is the lack of conditioning as exhibited in the fact

that independence holds iff $P(A|B) = P(A|B^c) = P(A)$ iff $P(B|A) = P(B|A^c) = P(B)$. The occurrence or nonoccurrence of B does not affect the likelihood of the occurrence of A, and the occurrence or nonoccurrence of A does not affect the likelihood of the occurrence of B.

Example B1-a

Consider two contractors working on two entirely different jobs. Let

A = event contractor "a" completes his job on schedule,

B = event contractor "b" completes his job on schedule.

It may well be that these two contractors work in a way that the performance of either has no affect on or relation to the performance of the other. Thus, it may be that $P(A|B) = P(A|B^c)$, in which case the common value is $P(A)$. We should thus assume, in modeling the situation, that $\{A,B\}$ is an independent pair of events. []

Suppose $\{A,B\}$ form an independent pair under the original probability measure. This independence is not an inherent property of the events (unless at least one is either the impossible event or the sure event). Stochastic independence is a property of the probability assignment, hence is determined by the probability measure $P(\cdot)$. Change to a new probability measure $P_1(\cdot)$ may destroy the stochastic independence. The following extension of the contractor example shows how stochastic independence may fail to hold, even though the contractors work "independently" in an operational sense. It also leads to the concept of conditional independence.

Example B1-b

Consider again the case of the two contractors. There may be some factor in the work situation which affects the performance of both. Suppose the jobs are outside, where performance can be affected by the weather. Let C = event the weather is "good". It may be reasonable to suppose that

$P(A|BC) = P(A|B^cC)$. That is, given good weather (i.e., the occurrence of C), the performance of contractor "b" has no effect on the performance of contractor "a". A similar situation may hold in the case of bad weather. Since $P(A|C) = P(AB|C) + P(AB^c|C) = P(B|C)P(A|BC) + P(B^c|C)P(A|B^cC)$, the equality $P(A|BC) = P(A|B^cC)$ implies that the common value is $P(A|C)$. Under these conditions, the pair $\{A,B\}$ will usually not be independent. There is a "probabilistic tie" between these two events by virtue of their relationships to the common event C. Let us examine the contractor example further by assigning some reasonable numerical values. Suppose

$$P(A|C) = 0.95 \qquad P(B|C) = 0.96 \qquad P(C) = 0.7$$

$$P(A|C^c) = 0.45 \qquad P(B|C^c) = 0.50 \qquad P(C^c) = 0.3.$$

Under the conditions $P(A|BC) = P(A|B^cC)$ and $P(A|BC^c) = P(A|B^cC^c)$, we have

$$P(AB) = P(C)P(B|C)P(A|BC) + P(C^c)P(B|C^c)P(A|BC^c)$$

$$= 0.7 \times 0.96 \times 0.95 + 0.3 \times 0.5 \times 0.45 = 0.7059$$

$$P(A)P(B) = [P(A|C)P(C) + P(A|C^c)P(C^c)]\,[P(B|C)P(C) + P(B|C^c)P(C^c)]$$

$$= [0.95 \times 0.7 + 0.45 \times 0.3]\,[0.96 \times 0.7 + 0.5 \times 0.3] = 0.6576.$$

Thus, $P(AB) \neq P(A)P(B)$, so $\{A,B\}$ is not independent. If the contractors work "independently", what is the tie between their performances? If A occurs, the likelihood of good weather is high, so that the likelihood of the occurrence of B is high. The numbers turn out to be $P(C|A) = P(A|C)P(C)/P(A) = 0.83 > 0.7 = P(C)$ and $P(B|A) = P(AB)/P(A) = 0.882 > 0.822 = P(B)$. If this is the only effective tie between events A and B, then once the weather is determined, there is no further influence of the performance of one contractor on that of the other. []

Let us examine further the assumption that $P(A|BC) = P(A|B^cC)$. Straightforward use of the defining relation for conditional probability and some elementary properties show that the following conditions are equivalent:

$$P(A|BC) = P(A|C) \quad P(B|AC) = P(B|C) \quad P(AB|C) = P(A|C)P(B|C)$$

$$P(A|B^cC) = P(A|C) \quad P(B^c|AC) = P(B^c|C) \quad P(AB^c|C) = P(A|C)P(B^c|C)$$

$$P(A^c|BC) = P(A^c|C) \quad P(B|A^cC) = P(B|C) \quad P(A^cB|C) = P(A^c|C)P(B|C)$$

$$P(A^c|B^cC) = P(A^c|C) \quad P(B^c|A^cC) = P(B^c|C) \quad P(A^cB^c|C) = P(A^c|C)P(B^c|C)$$

$$P(A|BC) = P(A|B^cC) \quad P(A^c|BC) = P(A^c|B^cC)$$

$$P(B|AC) = P(B|A^cC) \quad P(B^c|AC) = P(B^c|A^cC).$$

In view of our discussion above, it seems reasonable to call the common situation conditional independence, given C. Once C occurs, the occurrence or nonoccurrence of B does not further affect the likelihood of A, etc. As in the case of ordinary or total independence, we utilize the product rule as the basis of the mathematical definition, although some of the other equivalent relationships may be more useful in modeling.

DEFINITION. The pair $\{A,B\}$ of events is conditionally independent, given C, iff the product rule $P(AB|C) = P(A|C)P(B|C)$ holds.

An arbitrary class of events is conditionally independent, given C, iff a corresponding product rule holds for every finite subclass of two or more events from the class.

The product rule shows that conditional independence, given C, is just ordinary independence for the probability measure $P_C(\cdot) = P(\cdot|C)$. Conditioning by C leads to a new probability measure. In terms of this new probability measure, the pair $\{A,B\}$ is stochastically independent. As for the prior probability measure, we can assert that

> If any of the pairs $\{A,B\}$, $\{A,B^c\}$, $\{A^c,B\}$, or $\{A^c,B^c\}$ is conditionally independent, given C, then so are the others.

In Example B1-b, the conditioning event C is such that we have conditional independence, given C, and also, given C^c. If the weather

is good, the contractors work independently; they also work independently if the weather is bad. Such is not the case for all conditioning events.

Example B1-c

Suppose the two contractors of the previous example use some common item. Let D = event this item is in good supply and D^c = event this item is in short supply. If the supply is good, it is reasonable to suppose that the performance of one contractor has no effect on that of the other. Hence, it is reasonable to assume that $P(A|BD) = P(A|B^cD)$, which is equivalent to assuming $\{A,B\}$ is conditionally independent, given D. However, if the supply is short (i.e., if D^c occurs), the contractors may be in competition for the scarce item. Thus it may be reasonable to suppose $P(A|BD^c) < P(A|B^cD^c)$. If contractor "b" completes his job on time he has probably obtained the scarce item to the detriment of contractor "a". This condition violates one of the equivalent conditions for conditional independence of $\{A,B\}$, given D^c, so that we must assert conditional nonindependence. It is not difficult to show that in this case the pair $\{A,B\}$ is not totally independent. []

The following development shows that conditional independence, given one or both C and C^c, is unlikely to yield total independence. In the case of conditional independence, given C, and given C^c, we have

$$P(AB) = P(A|C)P(B|C)P(C) + P(A|C^c)P(B|C^c)P(C^c).$$

In the case of conditional independence, given C, but conditional nonindependence, given C^c, we have

$$P(AB) = P(A|C)P(B|C)P(C) + P(AB|C^c)P(C^c).$$

In either case, we have

$$P(A)P(B) = P(A|C)P(B|C)P^2(C) + P(A|C^c)P(B|C^c)P^2(C^c) + [P(A|C)P(B|C^c) + P(A|C^c)P(B|C)]P(C)P(C^c).$$

Only in unusual cases would we have $P(AB) = P(A)P(B)$. An example is provided in Problem B-5.

2. Some patterns of probable inference

We now consider a commonly encountered pattern of probable inference. We begin by giving two examples, then lifting out the essential pattern. When the appropriate conditional independence is identified, we show how it may help in determining the desired posterior odds.

Example B2-a

Associated with a certain disease are several symptoms. The presence of the symptoms does not guarantee the presence of the disease, but with high probability they occur when the disease is experienced and do not occur when the disease is absent. The symptoms are observed by chemical tests of blood samples. The tests themselves are not conclusive, but have high probability of detecting the presence or absence of the symptoms correctly. Now the chemical tests respond only to appropriate conditions in the blood and are not influenced by how the patient feels or otherwise responds to his condition. Let H = event the patient has the disease, D = event the symptoms occur (in the blood condition), and R = event the tests indicate the presence of the symptoms. Since the tests respond to the symptoms and not directly to the disease, it seems reasonable to suppose $P(R|DH) = P(R|DH^c)$ and $P(R|D^cH) = P(R|D^cH^c)$, so that $\{R,H\}$ is conditionally independent, given D, and given D^c. []

Example B2-b

A firm plans to market a new product nationally. Suppose the market may be characterized reasonably unambiguously as "favorable" or "unfavorable". The company executives decide to check market conditions in a test area. Let H = event the national market is favorable, D = event the test

market is favorable. Past experience allows reasonable estimates of $P(D|H)$ and $P(D|H^c)$. However, direct, completely reliable determination of the condition of the test-area market would be time consuming, expensive, and would entail the risk of a competitor capturing the market. A market survey of the test area is made. The results of such a survey are not conclusive; but under the assumed conditions, they are affected only by the conditions in the test area and not by existing conditions in the national market, except as the latter conditions are reflected in the test area. If R = event the survey shows the test market is favorable, we suppose that $P(R|DH) = P(R|DH^c)$ and $P(R|D^cH) = P(R|D^cH^c)$. This means that $\{R,H\}$ is conditionally independent, given D, and given D^c. []

These two examples exhibit features which are typical of a variety of inference problems.

1) There is an objective system about which some inference is to be made. In the first example, the objective system is the patient; in the second, it is the national market. The objective system is presumed to be in one of two objective states (the patient has the disease or does not; the market is favorable or is not). If H = event the objective system is in one of these states, then prior odds $P(H)/P(H^c) = a > 0$ are supposed known (or are estimated).

2) The objective system is not directly observable-- at least at the time of making the inference. But there is a data system which may be in one of several states (in each of the examples above, the data system is in one of two states). Each data state is "inconclusive" as to the objective state, but there is a "probabilistic linkage" between the data states and the objective states, expressed in terms of appropriate conditional probabilities, as follows. Let D_j = event the

data system is in state j (in the two-state system, we use D and D^c). We suppose the conditional probabilities $P(D_j|H) = b_j > 0$ and $P(D_j|H^c) = c_j > 0$ are known or may be estimated. Use of the ratio form of Bayes' rule shows the posterier odds to be $P(H|D_j)/P(H^c|D_j) = P(H)P(D_j|H)/P(H^c)P(D_j|H^c) = ab_j/c_j$.

3) In a typical situation, we do not have perfect information about the data state; rather, we have the report of an observer, or sensor. For simplicity, we discuss a two-state data system and let R = event the observer reports that D has occurred. If such a report is received, the effective posterior odds are $P(H|R)/P(H^c|R) = aP(R|H)/P(R|H^c)$. Since the objective system is not observable, $P(R|H)$ and $P(R|H^c)$ are usually not known. We suppose information is available about the reliability of the observer. That is, we suppose information is available to estimate $P(R|D) = d$ and $P(R|D^c) = e$, with $0 < d < 1$ and $0 < e < 1$. Note that "perfect information" about the data system requires $e = 0$ (for any positive value of d).

4) If the objective system is not observable, only the condition of the data system should affect the report. Thus, we should have $P(R|DH) = P(R|DH^c)$ and $P(R|D^cH) = P(R|D^cH^c)$. This is precisely the condition that $\{R,H\}$ is conditionally independent, given D, and given D^c. This does not imply that $\{R,H\}$ is independent.

5) Let us see how the assumption of conditional independence may help in determining the posterior odds, given the report.

$$\frac{P(H|R)}{P(H^c|R)} = \frac{P(H)}{P(H^c)} \cdot \frac{P(RD|H) + P(RD^c|H)}{P(RD|H^c) + P(RD^c|H^c)}$$

$$= a \cdot \frac{P(D|H)P(R|DH) + P(D^c|H)P(R|D^cH)}{P(D|H^c)P(R|DH^c) + P(D^c|H^c)P(R|D^cH^c)}.$$

Under the assumed conditional independence, this becomes

$$\frac{P(H\,|R)}{P(H^C|R)} = a \cdot \frac{P(D|H\,)P(R|D) + P(D^C|H\,)P(R|D^C)}{P(D|H^C)P(R|D) + P(D^C|H^C)P(R|D^C)}$$

which may be determined from the data available.

For a more general formulation of this problem, with more than two objective states and more than two data states, see Schum and Pfeiffer [1973]. To illustrate the analysis, we return to the previous examples.

Example B2-a (Continued)

The objective system is the patient, selected at random from among those who present themselves at the clinic, and H = event the patient has the disease in question. Suppose 10 percent of the patients examined at the clinic have the disease. Then prior odds $a = P(H)/P(H^C) = 1/9$. The data system is the blood condition. Let D = event the patient has the symptoms associated with the disease. Previous clinical experience shows $P(D|H) = 0.96$ and $P(D^C|H^C) = 0.95$. Let R = event the symptoms are indicated. The reliability of the testing procedure is such that $P(R|D) = 0.97$ and $P(R|D^C) = 0.01$. The patient is examined, a blood test made, and the report is found to be positive (i.e., event R occurs). According to the pattern above

$$\frac{P(H|R)}{P(H^C|R)} = (1/9)\,\frac{0.96 \times 0.97 + 0.04 \times 0.01}{0.05 \times 0.97 + 0.95 \times 0.01} = \frac{9316}{5220} \approx 1.78.$$

The positive result of the test changes the prior odds by a factor of about 16. The conditional probability that the patient has the disease, given the test result is $P(H|R) = \frac{9316/5220}{1 + 9316/5220} \approx 0.64$. []

We extend the second example to a slightly more general situation.

Example B2-b (Continued)

Consider the test-market problem described above. Initially, company executives think the odds for a favorable market are $P(H)/P(H^C) = 3$.

Past studies indicate $P(D|H) = 0.8$ and $P(D|H^c) = 0.2$. If the test market if found to be favorable (event D occurs), then

$$\frac{P(H|D)}{P(H^c|D)} = \frac{P(D|H)P(H)}{P(D|H^c)P(H^c)} = \frac{0.8}{0.2} \times 3 = 12$$

However, direct, completely reliable checking of even the test market conditions would be time consuming, expensive, and would entail the risk of a competitor capturing the market. Two market-survey firms are employed to survey the test market. Each makes a survey and reports its conclusion about the condition of the market. Let

A = event firm "a" reports the test market is favorable

B = event firm "b" reports the test market is favorable

The companies work "independently" in such a way that the investigation carried out by one does not affect that carried out by the other, regardless of the state of the test market. Because of the nature of the surveys, the results cannot be completely reliable. Suppose

$$P(A|D) = 0.9,\ P(A|D^c) = 0.3,\ P(B|D) = 0.8, \text{ and } P(B|D^c) = 0.2.$$

Find the posterior odds $P(H|AB)/P(H^c|AB)$ for a favorable market if both reports are favorable.

SOLUTION.

Again, we are faced with the problem of "independent tests." Complete independence of $\{A,B\}$ is not expected, for the outcomes of both tests are related to the condition of the test market. However, since the survey teams work in an operationally independent manner and neither team is affected by the national market except as it influences the test market, it seems reasonable to assume that $P(A|BD) = P(A|B^cD)$, $P(A|HD) = P(A|H^cD)$, $P(B|HD) = P(B|H^cD)$, and $P(AB|HD) = P(AB|H^cD)$. These conditions imply

that $\{A,B,H\}$ is conditionally independent, given D. A parallel argument yields conditional independence, given D^c. We note that

$$P(AB|H) = P(ABD|H) + P(ABD^c|H) = P(D|H)P(AB|DH) + P(D^c|H)P(AB|D^cH)$$

$$= P(D|H)P(A|D)P(B|D) + P(D^c|H)P(A|D^c)P(B|D^c)$$

and similarly for conditioning event H^c. We may, therefore, write

$$\frac{P(H|AB)}{P(H^c|AB)} = \frac{P(H)P(AB|H)}{P(H^c)P(AB|H^c)}$$

$$= \frac{P(H)}{P(H^c)} \frac{P(D|H)P(A|D)P(B|D) + P(D^c|H)P(A|D^c)P(B|D^c)}{P(D|H^c)P(A|D)P(B|D) + P(D^c|H^c)P(A|D^c)P(B|D^c)}$$

$$= 3 \frac{0.8\times0.9\times0.8 + 0.2\times0.3\times0.2}{0.2\times0.9\times0.8 + 0.8\times0.3\times0.2} = \frac{147}{16} \approx 9.2 .$$

The value 9.2 is somewhat less than the odds of 12 obtained if perfect information were available about the test market, as might be expected. []

If we do not have conditional independence, the problems are still meaningfull, but more detailed information is required for solution. Thus, we need $P(R|DH)$, $P(R|D^cH)$, $P(R|DH^c)$, and $P(R|D^cH^c)$. However, in this case it would be simpler to operate with $P(R|H)$ and $P(R|H^c)$, since R must be treated as a datum directly related to H. The reason for <u>not</u> doing this is that the objective system is not available for observation. But it is precisely in this situation that we should assume that $P(R|D) = P(R|DH)$, etc., since if the objective system is not available to the observer, only the condition of the data system can affect the report.

3. A classification problem

Suppose subjects are drawn from two groups. Each subject answers a battery of questions, or is otherwise tested with regard to a set of characteristics. The result is a profile of data for each subject tested. Each individual is to be classified in one of two groups, on the basis of the test results. The problem may be formulated in probabilistic terms as follows (see Schum and Pfeiffer [1977]).

There are n data classes $\mathcal{D}_i$, $i = 1, 2, \ldots, n$, one corresponding to each question or test. Let

D_{ij} = event the answer to question i falls into category (i,j)

Then $\mathcal{D}_i = \{D_{i1}, D_{i2}, \ldots, D_{im_i}\}$. If the list of possible answers or results is exhaustive and mutually exclusive, then $\mathcal{D}_i$ is a partition of the basic space on which probability is defined.

We suppose the subjects are drawn from two mutually exclusive groups. We let G_k = event the individual interviewed belongs to the kth group. In order to make probable inferences, we must suppose that the probabilities $P(G_k)$ and $P(D_{ij}|G_k)$ are positive and known. If we assume that no datum is conclusive, we must also have $0 < P(G_k|D_{ij}) < 1$ for all permissible i, j, k. Since each $\mathcal{D}_i$ is a partition, we have $\sum_j P(D_{ij}|G_k) = 1$ for each permissible i, k.

When an individual is interviewed, a profile is determined. A given profile corresponds to an event $E_p = D_{1j_1} D_{2j_2} \ldots D_{nj_n}$. The various possible profiles are mutually exclusive, so that events of the type E_p constitute a partition. We ask, "What is the inferential value of the compound event corresponding to a profile?" The usual answer is formulated in terms of the likelihood ratio $L_p = P(E_p|G_1)/P(E_p|G_2)$ or, equivalently, the log-likelihood ratio $\Lambda_p = \log L_p$. We may take logarithms to any base,

so long as we are consistent.

The problem, as it stands, would seem to require that we have conditional probabilities for each profile, for each of the two groups. This much data is rarely available, nor is it needed in a well-designed experiment. In the usual experimental design, an attempt is made to formulate the questions or tests in such a manner that responses or results are "independent." Once more, we have the issue of conditional independence. The probabilities of various answers to a given question should depend upon the basic characteristics of the subject (hence on his or her group membership), but should not depend upon his or her responses to the other questions . That is, a given subject's response to a particular question should be the same whether or not the other questions are asked, or regardless of the order in which they are asked. This does not mean that the responses to the questions are totally independent; the answers are conditioned by the group to which the subject belong (i.e., by the characteristics common to that group), else the questions have no diagnostic value. The desired independence holds within a given group, but the probability distributions are different in the two groups. Hence we make the assumption that the family $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ is <u>conditionally independent</u>, <u>given</u> G_1, and also <u>given</u> G_2. In this case

$$L_p = \prod_i L_{ij_i} = \prod_i \frac{P(D_{ij_i}|G_1)}{P(D_{ij_i}|G_2)} \qquad \text{and} \qquad \Lambda_p = \sum_i \Lambda_{ij_i} = \sum_i \log L_{ij_i}.$$

We may carry the formalism further in a useful way by introducing the random variables

$$T_i = \sum_j \Lambda_{ij} I_{D_{ij}} \qquad \text{(has value } \Lambda_{ij} \text{ whenever } D_{ij} \text{ occurs)}.$$

If $E_p = D_{1j_1} D_{2j_2} \cdots D_{nj_n}$ occurs, then $T = \sum_i T_i$ has the value $\Lambda_{1j_1} + \Lambda_{2j_2} + \cdots + \Lambda_{nj_n} = \Lambda_p$. Hence we utilize

$$T = \sum_i T_i = \sum_p \Lambda_p I_{E_p} \qquad \text{(has value } \Lambda_p \text{ whenever } E_p \text{ occurs)}.$$

Use of Bayes' theorem gives

$$\log \frac{P(G_1|E_p)}{P(G_2|E_p)} = \log \frac{P(E_p|G_1)P(G_1)}{P(E_p|G_2)P(G_2)} = T - t_c, \quad \text{where } t_c = \log \frac{P(G_2)}{P(G_1)}.$$

Standard practice is to classify the subject in group 1 iff $T > t_c$, which corresponds to $\dfrac{P(G_1|E_p)}{P(G_2|E_p)} > 1$.

This formulation allows us to deal with the problem of misclassification probabilities. Consider the conditional distribution functions $F_T(\cdot|G_1)$ and $F_T(\cdot|G_2)$, defined by $F_T(t|G_k) = P(T \leq t|G_k)$, $k = 1, 2$. In the conditionally independent case, $\{T_i: 1 \leq i \leq n\}$ is an independent class with respect to each of the probability measures $P(\cdot|G_1)$ and $P(\cdot|G_2)$. The central limit theorem ensures that for sufficiently large n both $F_T(\cdot|G_1)$ and $F_T(\cdot|G_2)$ are approximately normal. Examples show that the normal approximation may be quite useful for n as small as 4 or 5.

With the conditional distributions for T, standard statistical techniques may be utilized to determine the probabilities of misclassification errors. Under some conditions, better choices of the decision level t_c may be made. For a discussion of these issues, see Schum and Pfeiffer [1977].

Example B3-a

Subjects are to be classified in one of two groups. They are asked to

respond to a battery of six questions, each of which is to be answered in one of three ways: yes, no, uncertain. To calibrate the test, a sample of 100 subjects is interviewed intensively to determine the proper group classification for each. It is found that 55 belong in group 1 and 45 belong in group 2. If G_1 = event a subject belongs to group 1 and G_2 = event a subject belongs to group 2, these data are taken to mean that $P(G_1) = 0.55$ and $P(G_2) = 0.45$. The response of this control or calibration group to the questions is tabulated as follows:

Group 1 (55 members)

	Yes	No	Uncertain
j =	0	1	2
i = 1	17	26	12
2	7	30	18
3	8	40	7
4	14	31	10
5	15	25	15
6	9	33	13

n_{ij}

Group 2 (45 members)

	Yes	No	Uncertain
j =	0	1	2
i = 1	30	10	5
2	27	16	2
3	29	12	4
4	25	18	2
5	14	18	13
6	31	7	7

m_{ij}

We have assigned, arbitrarily, numbers 0, 1, 2 to the answers yes, no, uncertain, respectively. Thus, D_{10} is the event the answer to question 1 is "yes", D_{42} is the event the answer to question 4 is "uncertain," etc. We interpret the data in the tables to mean that $P(D_{10}|G_1) = n_{10}/55 = 17/55$ and $P(D_{42}|G_2) = m_{42}/45 = 2/45$, etc.

A subject is selected at random from the population from which the sample was taken. The subject's answers to the six questions, in order, are: yes, yes, no, uncertain, no, yes. How should this subject be classified?

SOLUTION.

The event $E_p = D_{10}D_{20}D_{31}D_{42}D_{51}D_{60}$ has occurred. We calculate the value

$T = \Lambda_p$ as follows:

$$\Lambda_{10} = \log P(D_{10}|G_1)/P(D_{10}|G_2) = \log \frac{17/55}{30/45} = -0.769$$

$$\Lambda_{20} = \log \frac{7/55}{27/45} = -1.551 \qquad \Lambda_{31} = \log \frac{40/55}{12/45} = 1.003$$

$$\Lambda_{42} = \log \frac{10/55}{2/45} = 1.409 \qquad \Lambda_{51} = \log \frac{25/55}{18/45} = 0.128$$

$$\Lambda_{60} = \log \frac{9/55}{31/45} = -1.437 \qquad \text{Summing gives } \Lambda_p = -1.217.$$

We also find $t_c = \log P(G_2)/P(G_1) = \log 0.45/0.55 = -0.201$. We thus have $T = \Lambda_p = -1.217 < -0.201 = t_c$; hence we classify the subject in group 2. To consider classification error probabilities, we could assume the conditional distributions for T, given G_1 and given G_2, to be approximately normal. By obtaining conditional means and variances for the various T_i, we could obtain the conditional means and variances for T, given G_1 and given G_2. Standard statistical methods could then be utilized. We do not pursue these matters, since our primary concern is the role of conditional independence in formulating the problem. []

It is not necessary that all the questions be conditionally independent. There could be some intentional redundancies, leading to conditional dependencies within each group. Suppose in the numerical example above that questions 1 and 2 were made to interlock. Then it would be necessary to consider this pair of questions as a single composite question with nine possible answers. Frequency data would be required on each pair of answers (no,no), (no, yes), (no, uncertain), (yes, no), (yes, yes), (yes, uncertain), (uncertain, no), (uncertain, yes), (uncertain, uncertain). One would still suppose conditional independence for the set of questions,

provided this composite question is dealt with as one question. More complex groupings could be made, increasing the amount of data needed to utilize the classification procedure, but there would be no difference in principle.

4. Problems

B-1 Prove the equivalence of at least four of the sixteen conditions for independence of $\{A,B\}$.

B-2 Complete the argument in Example B1-b to show that the equality $P(A|BC) = P(A|B^cC)$ implies that the common value is $P(A|C)$.

B-3 Establish the equivalence of at least four of the sixteen conditions for conditional independence of $\{A,B\}$, given C.

B-4 Show that the condition $P(A|BD^c) < P(A|B^cD^c)$ in Example B1-c implies $P(A|BD^c) < P(A|D^c)$.

B-5 A group of sixteen students has an equal number of males and females. One fourth of the females and three fourths of the males like to play basketball. One half of each likes to play volleyball. A student is selected from the group at random, on an equally likely basis. Let

A = event the student likes basketball,

B = event the student likes volleyball,

C = event the student is male.

Suppose $\{A,B\}$ is conditionally independent, given C, and conditionally independent, given C^c. Show that $\{A,B\}$ is independent and $\{B,C\}$ is independent, but $\{A,B,C\}$ is <u>not</u> independent.

B-6 In Example B2-b, show that the conditions i) $P(A|BD) = P(A|B^cD)$, ii) $P(A|HD) = P(A|H^cD)$, iii) $P(B|HD) = P(B|H^cD)$, and iv) $P(AB|HD) = P(AB|H^cD)$ together imply that $\{A,B,H\}$ is conditionally independent, given D.

B-7 In Example B2-b, determine $P(H|AB^c)/P(H^c|AB^c)$, the conditional odds, given conflicting reports of "favorable" by "a" and "unfavorable" by "b".

B-8 Consider the following problem, stated in a manner common in the literature. A patient is given a test for a type of cancer. The probability of a false positive is 0.10. The probability of a false negative is 0.20. One percent of the tested population is known to have the disease. If a patient receives two independent tests, and both are positive, find the probability the patient has cancer.

a) Let C = event the person selected has the given type of cancer

T_1 = event the first test indicates cancer (is positive),

T_2 = event the second test indicates cancer.

Discuss the reasonableness of the assumptions that $\{T_1, T_2\}$ is conditionally independent, given C, and is conditionally independent, given C^c.

b) Under these assumptions, determine $P(C|T_1T_2)$.

c) Under these assumptions, determine $P(C|T_1T_2^c)$.

B-9 A student decides to determine the odds on the forthcoming football game with State University. The odds depend heavily on whether State's star quarterback, recently injured, will play. A couple of phone calls yield two opinions whether the quarterback will play. Each report depends only on facts related to the condition of the quarterback and not on the outcome of the game (which is not known, of course). The two advisers have operated quite independently in arriving at their estimates. The student proceeds as follows. He lets

W = event the home team wins the game,

Q = event the star quarterback plays for State,

A = event the first informant is of the opinion he will play,

B = event the second informant is of the opinion he will play.

The student (having studied Example B2-b) decides to assume $\{W,A,B\}$ is conditionally independent, given Q, and conditionally independent,

given Q^c. On the basis of past experience he assesses the reliability of his advisers and assumes the following probabilities: $P(A|Q) = P(A^c|Q^c) = 0.8$, $P(B|Q) = 0.6$, and $P(B^c|Q^c) = 0.7$. Initially, he could only assume $P(Q) = P(Q^c) = 1/2$. Expert opinion assigns the odds $P(W|Q)/P(W^c|Q) = 1/3$ and $P(W|Q^c)/P(W^c|Q^c) = 3/2$. On the basis of these assumptions, determine the odds $P(W|AB^c)/P(W^c|AB^c)$ and the probability $P(W|AB^c)$.

B-10 A student is picked at random from a large freshman class in calculus. Let

T = event the student had a previous trigonometry course,

A = event the student made grade "A" on the first examination,

B = event the student made grade "B" or better in the course.

Data on the class indicate that

$P(T) = 0.60 \quad P(A|T) = 0.90 \quad P(A|T^c) = 0.30$

$P(B|AT) = P(B|A) = 0.60 \quad P(B|A^cT^c) = P(B|A^c) = 0.30.$

a) The student selected made "B" or better. What is the probability $P(T|B)$ that the student had a previous course in trigonometry?

b) Show that $\{T,B\}$ is not an independent pair.

B-11 Experience shows that 20 percent of the items produced on a production line are defective with respect to surface hardness. An inspection procedure has probability 0.1 of giving a false positive and probability 0.2 of giving a false negative. Units which fail to pass inspection are given a corrective treatment which has probability 0.95 of correcting any defective units and zero probability of producing any adverse effects on the essential properties of the units treated. However, with probability 0.3, the retreated units take on a characteristic color, regardless of whether or not they are defective (initially

or finally). Let

D_1 = event the unit selected is defective initially

I^c = event the unit failed inspection = event unit is retreated

D_2 = event the unit is defective after retreatment

C = event the unit is discolored after retreatment

a) Show that it is reasonable to suppose that $\{C, D_1\}$ is conditionally independent, given I^c, and that $\{C, D_2\}$ is conditionally independent, given $I^c D_1$. [Note that $IC = \emptyset$ and $P(D_1^c D_2) = 0$.]

b) Determine $P(D_2|C)$, the probability that a unit is defective, given that it is discolored.

B-12 In the classification problem, Example B3-a, determine the appropriate classification if the answers to the six questions are: yes, no, no, uncertain, yes, no, respectively.

C. Conditional Expectation

C. CONDITIONAL EXPECTATION

C. Conditional expectation

In order to introduce and develop the second concept of conditional independence, we need to examine the concept of conditional expectation. The usual introductory treatment of conditional expectation is intuitive, straightforward, but severely limited in scope. More general treatments tend to assume familiarity with advanced notions of measurability and abstract integration theory. We seek to bridge the gap and make the appropriate aspects of a general treatment more readily accessible.

1. Conditioning by an event.

If a conditioning event C occurs, we modify our probabilities by introducing the conditional probability measure $P(\cdot|C)$. Thus, $P(A)$ is replaced by $P(A|C) = P(AC)/P(C)$. In making this change, we do two things:

i) We limit the possible outcomes to those in event C

ii) We "normalize" the probability mass in C to make it the new unit of mass

It seems reasonable to make a corresponding modification of mathematical expectation, which we view as a probability weighted average of the values taken on by a random variable. Two possibilities are apparent.

a) We could modify the prior probability measure $P(\cdot)$ to the conditional probability measure $P(\cdot|C)$, then take expectation (i.e., weighted average) with respect to this new probability mass assignment.

b) We could continue to use the original probability measure $P(\cdot)$ and modify our averaging process as follows:

   i) For a real random variable X, we consider the value $X(\omega)$ for only those ω in the event C. We do this by utilizing the random variable $I_C X$, which has the value $X(\omega)$ for ω in C, and has the value zero for any ω outside C. Then $E[I_C X]$ is

the probability weighted sum of the values taken on by X in the event C.

ii) We divide the weighted sum by $P(C)$ to obtain the weighted average.

As shown by Theorem C1-1, below, these two approaches are equivalent. For reasons which will become more apparent in subsequent developments, we take the second approach as the basis for definition. For one thing, we can do the "summing" in each case with the prior probability measure, then obtain the average by dividing by $P(C)$ for the particular conditioning event. This approach facilitates relating the present concept to the more general concept of conditional expectation, given a random vector, which is developed in the next two sections.

DEFINITION. If the event C has positive probability and indicator function I_C, the conditional expectation of X, given C, is the quantity $E[X|C] = E[I_C X]/P(C)$.

Several properties may be established easily.

Theorem C1-1

a) $E[X|C]$ is expectation with respect to the conditional probability measure $P(\cdot|C)$

b) $E[I_A|C] = P(A|C)$

c) If $C = \bigcup_i C_i$ (disjoint union), then $E[X|C]P(C) = \sum_i E[X|C_i]P(C_i)$.

PROOF OF a)

If X is a simple random variable $\sum_k t_k I_{A_k}$, then

$$E[X|C] = E[I_C X]/P(C) = E[\sum_k t_k I_C I_{A_k}]/P(C) = \sum_k t_k E[I_{A_k C}]/P(C)$$
$$= \sum_k t_k P(A_k|C) = E_C[X]$$

where the symbol $E_C[\cdot]$ indicates expectation with respect to the conditional probability measure $P(\cdot|C)$.

If $X \geq 0$, then there is a sequence $\{X_n : 1 \leq n\}$ of simple random variables increasing to X. This ensures that the sequence $\{I_C X_n : 1 \leq n\}$ is a sequence of simple random variables increasing to $I_C X$. By definition,

$$E[I_C X]/P(C) = \lim_n E[I_C X_n]/P(C) \quad \text{and} \quad E_C[X] = \lim_n E_C[X_n].$$

Since $E[I_C X_n]P(C) = E_C[X_n]$ for each n, the limits must be the same. In the general case, we consider $X = X_+ - X_-$, with both $X_+ \geq 0$ and $X_- \geq 0$. By linearity,

$$E[I_C X]/P(C) = E[I_C X_+]/P(C) - E[I_C X_-]/P(C) = E_C[X_+] - E_C[X_-] = E_C[X].$$

Propositions b) and c) are established easily from properties of mathematical expectation. []

The following theorem provides a link between the present concept and the more general concept developed in the next two sections.

Theorem C1-2

If event $C = Y^{-1}(M) = \{Y \in M\}$, for any Borel set M, has positive probability, then $E[I_M(Y)g(X)] = E[g(X)|Y \in M]P(Y \in M)$.

PROOF

By Theorem A1-1, $I_M(Y) = I_C$. By definition $E[I_C g(X)] = E[g(X)|C]P(C)$. Hence, $E[I_M(Y)g(X)] = E[g(X)|Y \in M]P(Y \in M)$. []

It should be noted that both X and Y can be vector-valued. The function g must be real-valued, and M is any Borel set on the codomain of Y.

2. Conditioning by a random vector-- special cases

In this section, we consider two simple, but important, cases of conditional expectation, given a random vector. We make an intuitive approach, based on the idea of a conditional distribution. In each case, the conditional expectation is found to be of the form $E[g(X)|Y = u] = e(u)$, where $e(\cdot)$ is a Borel function defined on the range of Y. This function satisfies, in each case, a fundamental equation which provides a tie with the concept of conditional expectation, given an event, and which serves as the basis for a number of important properties. This fundamental equation also provides the basis for extending the concept of conditional expectation, given a random vector, to the general case.

Case i) X, Y discrete. $X = \sum_{i=1}^{n} t_i I_{A_i}$ and $Y = \sum_{j=1}^{m} u_j I_{B_j}$, where $A_i = \{\omega: X(\omega) = t_i\}$ and $B_j = \{\omega: Y(\omega) = u_j\}$. We suppose $P(A_i) > 0$ and $P(B_j) > 0$ for each permissible i,j. Now

$$
\begin{aligned}
E[g(X)|Y = u_j]P(Y = u_j) &= E[g(X)|B_j]P(B_j) \\
&= E[g(X)I_{B_j}] && \text{by def.} \\
&= E[g(X)I_{\{u_j\}}(Y)] && \text{by Thm A1-1} \\
&= \sum_{i,k} g(t_i)I_{\{u_j\}}(u_k)p_{XY}(t_i,u_k) \\
&= \sum_{i} g(t_i)p_{XY}(t_i,u_j) && \text{since } I_{\{u_j\}}(u_k) = 1 \\
& && \text{iff } j = k.
\end{aligned}
$$

If we consider the conditional probability mass function

$$
p_{X|Y}(t_i|u_j) = \frac{p_{XY}(t_i,u_j)}{p_Y(u_j)} = \frac{P(X = t_i,\ Y = u_j)}{P(Y = u_j)}
$$

we may write

$$
E[g(X)|Y = u_j]P(Y = u_j) = \left[\sum_i g(t_i)p_{X|Y}(t_i|u_j)\right]p_Y(u_j)
$$

from which we get

$$
E[g(X)|Y = u_j] = \sum_i g(t_i)p_{X|Y}(t_i|u_j) = e(u_j) \quad \text{for each } u_j \text{ in the range of } Y.
$$

We may let $e(\cdot)$ be any continuous function which takes on the prescribed values $e(u_j)$ for each u_j in the range of Y. Then $e(\cdot)$ is a Borel function. Suppose M is any Borel set on the codomain of Y. Then

$$\begin{aligned} E[I_M(Y)g(X)] &= \sum_{i,k} g(t_i)I_M(u_k)p_{XY}(t_i,u_k) \\ &= \sum_{i,k} g(t_i)I_M(u_k)p_{X|Y}(t_i|u_k)p_Y(u_k) \\ &= \sum_k I_M(u_k)\Big[\sum_i g(t_i)p_{X|Y}(t_i|u_k)\Big]p_Y(u_k) \\ &= \sum_k I_M(u_k)e(u_k)p_Y(u_k) = E[I_M(Y)e(Y)]. \end{aligned}$$

Hence, $e(\cdot)$ must satisfy

$$E[I_M(Y)g(X)] = E[I_M(Y)e(Y)] \quad \forall \text{ Borel set } M \text{ in the codomain of } Y$$

The uniqueness property E7) for expectations ensures $e(\cdot)$ is unique a.s. $[P_Y]$, which in this case means $e(\cdot)$ is uniquely determined on the range of Y. []

Example C2-a

Suppose X, Y produce the joint distribution shown in Fig. C2-1. Determine the function $e(\cdot) = E[X|Y = \cdot]$.

SOLUTION.

From the joint distribution, we obtain the quantities

$$p_Y(1) = p_Y(2) = 3/10 \qquad p_Y(3) = 4/10$$

$$p_{X|Y}(1|1) = p_{X|Y}(2|1) = p_{X|Y}(3|1) = \frac{1/10}{3/10} = 1/3$$

$$p_{X|Y}(4|1) = p_{X|Y}(5|1) = 0.$$

Hence $e(1) = 1/3(1 + 2 + 3) = 2$.

Similarly $e(2) = 1/3(2 + 3 + 4) = 3$ and $e(3) = 1/4(2 + 3 + 4 + 5) = 7/2$

Graphical interpretation. The conditional probabilities $p_{X|Y}(k|u)$, for fixed u, are proportional to the probability masses on the horizontal line

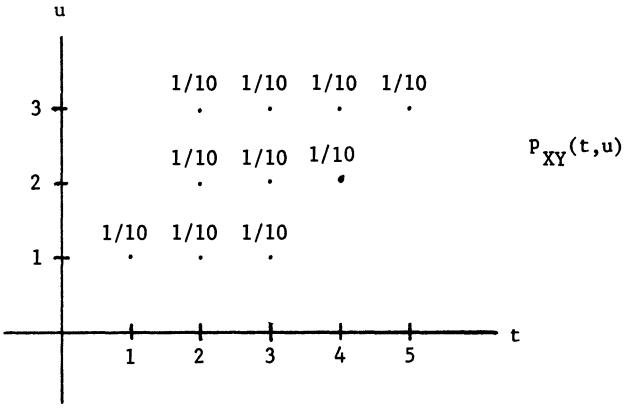


Figure C2-1. Joint distribution for Example C2-a.

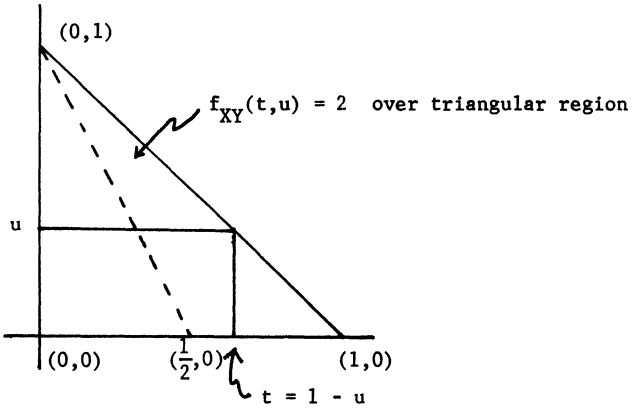


Figure C2-2. Joint distribution for Example C2-b.

corresponding to $Y = u$. Thus, $E[X|Y = u]$ is the center of mass for that part of the joint distribution which corresponds to $Y = u$. []

Case ii) X, Y are absolutely continuous, with joint density function f_{XY}. Since the event $\{Y = u\}$ has zero probability, we cannot begin with conditional expectation, given the event $\{Y = u\}$. We may utilize the intuitive notion of a conditional distribution, given $Y = u$, by employing the following device. Let

$$f_{X|Y}(t|u) = \begin{cases} f_{XY}(t,u)/f_Y(u) & \text{for } f_Y(u) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For fixed u such that $f_Y(u) > 0$ (i.e., in the range of Y), the function $f_{X|Y}(\cdot|u)$ has the properties of a density function: $f_{X|Y}(t|u) \geq 0$ and $\int f_{X|Y}(t|u)\, dt = 1$. It is natural to call this the conditional density function for X, given $Y = u$. In part, the terminology is justified by the following development. Let M be any Borel set on the codomain of Y. Then

$$\begin{aligned} E[g(X)I_M(Y)] &= \iint g(t)I_M(u)f_{XY}(t,u)\, dtdu \\ &= \int I_M(u)\left[\int g(t)f_{X|Y}(t|u)\, dt\right] f_Y(u)\, du \\ &= \int I_M(u)e(u)f_Y(u)\, du = E[I_M(Y)e(Y)] \end{aligned}$$

where $e(u) = \int g(t)f_{X|Y}(t|u)\, dt$.

Now $e(\cdot)$ must satisfy

$$E[I_M(Y)g(X)] = E[I_M(Y)e(Y)] \quad \forall \text{ Borel sets } M \text{ in the codomain of } Y.$$

It seems natural to call $e(u)$ the conditional expectation of $g(X)$, given $Y = u$. In the case $P(Y \in M) > 0$, we have by Theorem C1-2

$$E[I_M(Y)e(Y)] = \int I_M(u)e(u)f_Y(u)\, du = E[g(X)|Y \in M]P(Y \in M).$$

If $e(\cdot)$ is Borel, as it will be in any practical case, property E7) for expectation ensures that $e(Y)$ is a.s. unique, or $e(\cdot)$ is unique a.s. $[P_Y]$, which means that it is determined essentially on the range of Y. []

Example C2-b

Suppose X, Y produce a joint distribution which is uniform over the triangular region with vertices (0,0), (1,0), (0,1), as shown in Fig. C2-2. Now

$$f_Y(u) = \int f_{XY}(t,u)\, dt = 2\int_0^{1-u} dt = 2(1 - u) \qquad 0 \le u \le 1$$

and

$$f_{X|Y}(t,u) = \frac{1}{1 - u} \quad \text{for} \quad 0 \le t \le 1 - u, \quad 0 \le u < 1 \quad \text{(and zero elsewhere).}$$

Hence

$$e(u) = E[X|Y = u] = \int t\, f_{X|Y}(t|u)\, dt = \frac{1}{1 - u}\int_0^{1-u} t\, dt = \frac{1 - u}{2} \qquad 0 \le u < 1.$$

Graphical interpretation. The dashed line in Fig. C2-2 is the graph of e(u) vs. u. This could have been anticipated by the following graphical interpretation. If f_{XY} is continuous, we may visualize $f_{X|Y}(t|u)$ as proportional to the mass per unit length in a very narrow strip on the plane about the line corresponding to $Y = u$. $E[X|Y = u]$ is the center of mass of the portion of the joint distribution lying in that narrow strip. []

3. Conditioning by a random vector-- general case

The treatment of the special cases in the previous section begins with the notion of a conditional distribution. While this approach is intuitively appealing, and quite adequate for the simplest cases, it quickly becomes unmanageable in more general cases which involve random vectors of higher dimensions with mixed distributions. We seek a more satisfactory approach.

We base our development on a simple property derived in each of the two special cases considered in the previous section. In each case, the quantity called the conditional expectation of $g(X)$, given $Y = u$, is the value $e(u)$ of a Borel function $e(\cdot)$ which is defined on the range of Y. The random variable $e(Y)$ satisfies

A) $E[I_M(Y)g(X)] = E[I_M(Y)e(Y)]$ $\forall$ Borel sets M in the codomain of Y.

By the uniqueness property E6) for mathematical expectation, $e(Y)$ must be a.s. unique, which is equivalent to the condition $e(\cdot)$ is unique a.s. $[P_Y]$. By Theorem C1-2 on conditional expectation, given an event, we have

B) If $P(Y \in M) > 0$, then $E[I_M(Y)e(Y)] = E[g(X)|Y \in M]P(Y \in M)$.

Motivated by these developments, we make the

DEFINITION. Let $e(\cdot)$ be a real-valued, Borel function defined on a set which includes the range of random vector Y. Then the quantity $e(u)$ is the conditional expectation of $g(X)$, given $Y = u$, denoted $E[g(X)|Y = u]$ iff

A) $E[I_M(Y)g(X)] = E[I_M(Y)e(Y)]$ $\forall$ Borel sets M in the codomain of Y.

Associated with the Borel function $e(\cdot)$ is the random variable $e(Y)$. Now $e(\cdot)$ is unique a.s. $[P_Y]$ and $e(Y)$ is unique a.s.

DEFINITION. The random variable $e(Y)$ is called the conditional expectation of $g(X)$, given Y, denoted $E[g(X)|Y]$.

Note that we must distinguish between the two symbols:

a) $E[g(X)|Y = \cdot] = e(\cdot)$, a Borel function on the range of Y

b) $E[g(X)|Y] = e(Y)$ a random variable-- for a given ω we write $E[g(X)|Y](\omega)$.

Example C3-a

If the conditioning random vector Y is simple, an explicit representation of $e(Y) = E[g(X)|Y]$ is obtained easily. Suppose $Y = \sum_{j=1}^{m} u_j I_{B_j}$ (in canonical form-- see Sec A1), so that $B_j = \{Y = u_j\}$ and $I_{B_j} = I_{\{u_j\}}(Y)$. If $e(u) = E[g(X)|Y = u]$, then $e(\cdot)$ is defined for u_j in the range of Y by $e(u_j) = E[g(X)|Y = u_j] = E[I_{\{u_j\}}(Y)g(X)]/P(Y = u_j)$ (conditional expectation, given the event $\{Y = u_j\}$). Hence,

$$e(Y) = \sum_{j=1}^{m} e(u_j) I_{B_j} = \sum_{j=1}^{m} E[g(X)|Y = u_j] I_{\{u_j\}}(Y).$$

Thus, when the conditioning random vector is simple, so that $P(Y = u_j) > 0$, the concepts of conditional expectation, given the event $\{Y = u_j\}$, and of conditional expectation, given $Y = u_j$, coincide for u_j in the range of Y, and the same symbol is used for both. Use of formula B), above, gives

$$E[g(X)|Y \in M]P(Y \in M) = E[I_M(Y)e(Y)] = \sum_{j=1}^{m} E[g(X)|Y = u_j]E[I_M(Y)I_{\{u_j\}}(Y)].$$

The quantity $E[I_M(Y)I_{\{u_j\}}(Y)] = P(Y = u_j)$ iff $u_j \in M$, and is zero otherwise. []

Example C3-b

Consider the random variables X, Y in Example C2-b. Let M be the semi-infinite interval $(-\infty, 0.5]$, so that $\{Y \in M\} = \{Y \leq 0.5\}$. Then

$$P(Y \in M) = \int_{-\infty}^{0.5} f_Y(u)\, du = 3/4 \quad \text{[May be obtained geometrically.]}$$

$$E[I_M(Y)e(Y)] = \int_{-\infty}^{0.5} e(u) f_Y(u)\, du = \int_0^{0.5} (1 - u)^2\, du = 7/24.$$

Hence

$$E[X|Y \leq 0.5] = (7/24)/(3/4) = 7/18. \qquad []$$

In each of the two special cases considered in Sec C2, we have been able to produce a Borel function $e(\cdot)$ which satisfies the defining relation A) for conditional expectation. The uniqueness property E6) shows $e(\cdot)$ to be unique a.s. $[P_Y]$. In Sec C4, we state a number of properties of conditional expectation which provide the basis for much of its usefulness. In Sec C7, we provide proofs of these properties based on proposition A) and properties E1) through E6) for expectation. These properties hold whenever the appropriate Borel function $e(\cdot)$ exists. Thus, they hold for the two special cases examined in Sec C2 and for others which can be derived similarly. It would be convenient if we knew the conditions under which suitable $e(\cdot)$ exists. As a matter of fact, if we utilize the powerful existence theorem E10) for mathematical expectation, stated without proof in Sec A2, we may assert the existence of $e(\cdot)$ for any random vectors X, Y and any real-valued Borel function $g(\cdot)$ such that $E[g(X)]$ is finite. The properties obtained in Sec C7 then hold in any such case.

4. Properties of conditional expectation

In this section, we list the principal properties of conditional expectation, given a random vector, which are utilized in subsequent developments. Proofs are given in Sec C7. These are based on the defining relation A) and properties E1) through E6) for mathematical expectation.

In the following, we suppose, without repeated assertion, that the random vectors and Borel functions are such that the existence of ordinary expectations is assured.

We begin our list of properties with the defining condition.

CE1) $e(Y) = E[g(X)|Y]$ a.s. iff $E[I_M(Y)e(Y)] = E[I_M(Y)g(X)]$ for all Borel sets M in the codomain of Y.

As noted in relation B), in Sec C3,

CE1a) If $P(Y \in M) > 0$, then $E[I_M(Y)e(Y)] = E[g(X)|Y \in M]P(Y \in M)$.

If, in CE1), we let M be the entire codomain of Y, so that $I_M(Y)$ has the constant value one for all ω, we obtain the important special case

CE1b) $E[g(X)] = E\{E[g(X)|Y]\}$.

The device of first conditioning by a random vector Y and then taking expectations is often useful, both in applications and in theoretical developments. As a simple illustration of the process, we continue an earlier example.

Example C4-a (Continuation of Example C2-b)

Consider, again, the random variables X,Y which produce a joint distribution which is uniform over the triangular region with vertices $(0,0)$, $(1,0)$, $(0,1)$. It is shown in Example C2-b that

$$f_Y(u) = 2(1 - u) \quad \text{for} \quad 0 \leq u \leq 1 \quad \text{(and zero elsewhere)}$$

$$e(u) = E[X|Y = u] = \frac{1 - u}{2} \quad \text{for} \quad 0 \leq u < 1.$$

By CE1b)

$$E[X] = E[e(Y)] = \int e(u) f_Y(u)\, du = \int_0^1 (1 - u)^2\, du = 1/3.$$

The result could, of course, have been obtained by finding $f_X(t) = 2(1 - t)$ for $0 \leq t \leq 1$ and calculating

$$E[X] = \int t f_X(t)\, dt = 2 \int_0^1 (t - t^2)\, dt = 1/3.$$

The choice of approach depends upond the objectives and the information at hand. []

The next three properties emphasize the integral character of conditional expectation, since they are in direct parallel with basic properties of expectations or integrals. One must be aware, of course, that for conditional expectation the properties may fail to hold on an exceptional set of outcomes whose probability is zero. The proofs given in Sec C7 show how these properties are, in fact, based on corresponding properties of mathematical expectation.

CE2) Linearity. $E[ag(X) + bh(Y)|Z] = aE[g(X)|Z] + bE[h(Y)|Z]$ a.s. (with extension by mathematical induction to any finite linear combination.)

CE3) Positivity; monotonicity.

$g(X) \geq 0$ a.s. implies $E[g(X)|Y] \geq 0$ a.s.

$g(X) \geq h(Y)$ a.s. implies $E[g(X)|Z] \geq E[h(Y)|Z]$ a.s.

CE4) Monotone convergence. $X_n \to X$ a.s. monotonically implies $E[X_n|Y] \to E[X|Y]$ a.s. monotonically

Independence of random vectors is associated with a lack of "conditioning" in the following sense.

CE5) Independence. The pair $\{X,Y\}$ is independent iffi $E[g(X)|Y] = E[g(X)]$ a.s. for all Borel functions g such that $E[g(X)]$ is finite, iffi $E[I_N(X)|Y] = E[I_N(X)]$ a.s. for all Borel sets N on the codomain of X.

Note that it is not sufficient that $E[g(X)|Y] = E[g(X)]$ a.s. for one specific Borel function g. It is relatively easy to establish counter-examples (see Problem C-5).

Use of linearity, monotone convergence, and approximation of Borel functions by step functions (simple functions) yields an extension of CE1).

CE6) $e(Y) = E[g(X)|Y]$ a.s. iff $E[h(Y)g(X)] = E[h(Y)e(Y)]$ for all Borel functions h such that the expectations exist.

The next three properties exhibit distinctive features of conditional expectation which are the basis of much of their utility. Proofs rest on previously established properties of mathematical expectation, especially part a) of E6). We employ these properties repeatedly in subsequent developments.

CE7) If $X = h(Y)$, then $E[g(X)|Y] = g(X)$ a.s.

CE8) $E[h(Y)g(X)|Y] = h(Y)E[g(X)|Y]$ a.s.

CE9) If $Y = h(W)$, then $E\{E[g(X)|Y]|W\} = E\{E[g(X)|W]|Y\} = E[g(X)|Y]$ a.s.

It occurs frequently that Y is a random vector whose coordinates form a subset of the coordinates of W. Thus, we may consider $W = (Y,Z)$, which implies Y is a Borel function of W, so that

CE9a) $E\{E[g(X)|Y]|Y,Z\} = E\{E[g(X)|Y,Z]|Y\} = E[g(X)|Y]$ a.s.

If the function h in CE9) has a Borel inverse, then $W = h^{-1}(Y)$, so that the roles of Y and W are interchangeable. Thus, we may assert

CE9b) If $Y = h(W)$, where h is Borel with a Borel inverse,

then $E[g(X)|Y] = E[g(X)|W]$ a.s.

We note two special cases of CE9b). If the coordinates of Y are obtained as a permutation of the coordinates of W, then $Y = h(W)$, where h is one-one, onto, and continuous, hence Borel with Borel inverse. Thus, conditioning by a random vector does not depend upon the particular ordering of the coordinates. If we have a pair of random vectors $\{X,Y\}$ which do

not share any coordinates, then conditioning by the pair is understood as conditioning by the random vector (X,Y) whose coordinates consist of the combined set of coordinates of the two random vectors (in any order). In a similar manner, we can consider two random vectors which may have some coordinates in common. Conditioning by such a pair is understood as conditioning by a random vector whose coordinates consist of the combined set of distinct coordinates. For example, suppose $X = (X_1,X_2,X_3)$ and $Y = (X_1,X_3,X_4)$. Then conditioning by X,Y is conditioning by $W = (X_1,X_2,X_3,X_4)$. It is apparent how these ideas extend to larger combinations of vectors.

The next result is so plausible that it is frequently taken to be self evident. Although it is easily established in certain simple cases, it is somewhat difficult to establish in the general case, as noted in Sec C7. It is extremely useful in the Borel function form, as follows.

CE10) Suppose g is a Borel function such that $E[g(X,v)]$ is finite for all v in the range of Y and $E[g(X,Y)]$ is finite. Then $E[g(X,Y)|Y = u] = E[g(X,u)|Y = u]$ a.s. $[P_Y]$.

In the independent case, CE10) takes a useful form.

CE11) If the pair $\{X,Y\}$ in CE10) is independent, then $E[g(X,Y)|Y = u] = E[g(X,u)]$ a.s.

Among the inequalities for expectations which can be extended to conditional expectations, the following are useful in many applications.

CE12) Triangle inequality. $|E[g(X)|Y]| \leq E[|g(X)||Y]$ a.s.

CE13) Jensen's inequality. If g is a convex function on an interval I which contains the range of real random variable X, the $g(E[X|Y]) \leq E[g(X)|Y]$ a.s.

Establishment of inequalities for conditional expectation (as for expectation)

depends upon setting up the appropriate inequalities for random variables, then utilizing monotonicity CE3). The inequalities on the random variables are often expressions of classical inequalities in ordinary analysis. As in the case of expectations, monotone convergence plays a key role in establishing analogs of Fatou's lemma, dominated convergence, and countable additivity.

*5. Conditional distributions

The introductory treatment of the special cases of conditional expectation in Sec C2 utilizes the notion of conditional distribution. In Sec C3, however, we disregard this notion in developing the general concept of conditional expectation, given a random vector. In the present section, we show that conditional probability and conditional distributions can be treated as special cases of conditional expectation.

By properties CE1b) and E1a),

$$E\{I_M(Y)E[I_N(X)|Y]\} = \int_M E[I_N(X)|Y = u]\, dF_Y(u)$$
$$= E[I_N(X)|Y \in M]P(Y \in M) = P(X \in N|Y \in M)P(Y \in M).$$

This leads naturally to the

DEFINITION. $P(X \in N|Y = u) = E[I_N(X)|Y = u]$ a.s.

If X is real-valued and $N_t = (-\infty, t]$, then we set

$$F_{X|Y}(t|u) = P(X \leq t|Y = u) = E[I_{N_t}(X)|Y = u] \quad \text{a.s.}$$

For each fixed t, this defines a Borel function of u with properties which suggest that for each fixed u in the range of Y the function $F_{X|Y}(\cdot|u)$ should be a distribution function. One property of interest is the following.

$$P(X \leq t, Y \in M) = E[I_{N_t}(X)I_M(Y)] = E\{I_M(Y)E[I_{N_t}(X)|Y]\}$$
$$= \int_M F_{X|Y}(t|u)\, dF_Y(u)$$

from which it follows as a special case that

$$F_X(t) = E\{E[I_{N_t}(X)|Y]\} = \int F_{X|Y}(t|u)\, dF_Y(u).$$

This last equality is often known as the law of total probability, since it appears as a generalization of a rule known by that name,

$$P(A) = \sum_i P(A|B_i)P(B_i), \quad \text{where} \quad A \subset \biguplus_i B_i.$$

---

* The material in this section is not needed in the subsequent sections and may be omitted without loss of continuity.

There are some technical difficulties in dealing with $F_{X|Y}(\cdot|u)$ as a distribution function. These arise because for each real t there is an exceptional set of u of P_Y measure zero. That is $F_{X|Y}(t|u) = P(X \leq t|Y = u)$ a.s. $[P_Y]$. Since there is an uncountable infinity of real numbers t, certain problems can arise with which we are not equipped to deal. In the case of joint density functions or of jointly discrete random variables, the motivating treatment of Sec C2 indicates the problem may be solved. For a real random variable X, a distribution function is determined by its values on the rationals, which involves only a countable infinity of values. Thus, it is known that for real random variable X and any random vector Y there is a <u>regular conditional distribution function, given</u> Y, with the properties

1) $F_{X|Y}(\cdot|u)$ is a distribution function for a.e. u $[P_Y]$,

2) For each real t, $F_{X|Y}(t|u) = P(X \leq t|Y = u)$ for a.e. u $[P_Y]$,

3) $E[g(X)|Y = u] = \int g(t)\, dF_{X|Y}(t|u)$ for a.e. u $[P_Y]$.

In some cases, for a.e. fixed u, $F_{X|Y}(\cdot|u)$ is differentiable and the function $f_{X|Y}(\cdot|u)$ defined by

$$f_{X|Y}(t|u) = \frac{d}{dt} F_{X|Y}(t|u)$$

is a <u>conditional density function for</u> X, <u>given</u> $Y = u$. This agrees with the conditional density function introduced in Sec C2.

As an important example of the use of these ideas, consider the problem of determining the distribution for the sum $Z = X + Y$ of two random variables X,Y. If we let $Q = \{(t,u): t + u \leq v\}$ (see Fig. C5-1), then

$$F_Z(v) = P(X + Y \leq v) = P[(X,Y) \in Q] = E[I_Q(X,Y)] = E\{E[I_Q(X,Y)|Y]\}$$
$$= \int E[I_Q(X,u)|Y = u]\, dF_Y(u) \qquad \text{by CE10).}$$

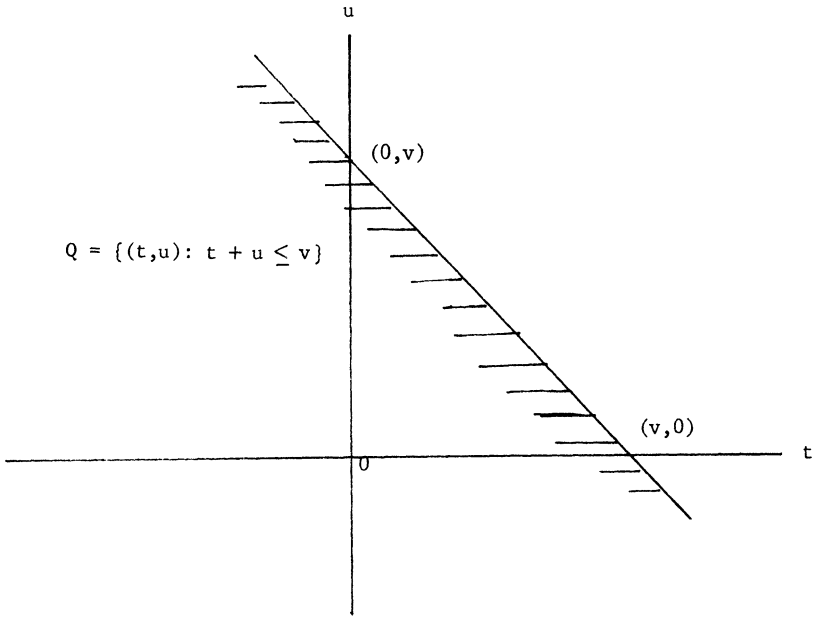


Figure C5-1. The region Q for Z = X + Y.

For each fixed v, $(t,u) \in Q$ iff $t + u \leq v$ iff $t \leq v - u$ iff $t \in N_{v-u}$. Hence, $I_Q(X,u) = I_{N_{v-u}}(X)$, so that

$$E[I_Q(X,u) \mid Y = u] = F_{X|Y}(v-u|u).$$

If $F_{X|Y}$ is a regular conditional distribution, then

$$F_Z(v) = \int F_{X|Y}(v - u|u)\, dF_Y(u).$$

If $\{X,Y\}$ is an independent pair, then $F_{X|Y}(v - u|u) = F_X(v - u)$ a.s. $[P_Y]$, so that

$$F_Z(v) = \int F_X(v - u)\, dF_Y(u).$$

This last combination is known as the <u>convolution</u> of F_X with F_Y.

6. Conditional distributions and Bayes' theorem

We suppose a regular conditional distribution has been determined. It frequently is necessary to reverse the conditioning, as in Bayes theorem for events. In the following we treat X, Y as real-valued. Extensions to the vector-valued cases are immediate.

a) If both X,Y are discrete, there is no problem. If we let $p_{X|Y}(t_i|u_j) = P(X = t_i|Y = u_j)$, and similarly for the other cases, then

$$p_{Y|X}(u_j|t_i) = \frac{p_{X|Y}(t_i|u_j)\, p_Y(u_j)}{p_X(t_i)},$$

b) If there is a joint density function, then by definition

$$f_{Y|X}(u|t) = \frac{f_{X|Y}(t|u)f_Y(u)}{f_X(t)} \qquad \text{for} \quad f_X(t) > 0,$$

c) Suppose X is discrete and Y is absolutely continuous.

$$\begin{aligned} F_{Y|X}(u|t_i) &= P(Y \le u|X = t_i) \\ &= E[I_{N_u}(Y)I_{\{t_i\}}(X)]/E[I_{\{t_i\}}(X)] \\ &= \frac{E\{I_{N_u}(Y)E[I_{\{t_i\}}(X)|Y]\}}{E\{E[I_{\{t_i\}}(X)|Y]\}} \qquad \text{by CE1)} \\ &= \frac{\int_{-\infty}^{u} P(X = t_i|Y = v)f_Y(v)\, dv}{\int P(X = t_i|Y = v)f_Y(v)\, dv}. \end{aligned}$$

Differentiation by u gives

$$f_{Y|X}(u|t_i) = \frac{P(X = t_i|Y = u)f_Y(u)}{P(X = t_i)}.$$

Simple algebraic manipulation gives

$$P(X = t_i|Y = u) = \frac{f_{Y|X}(u|t_i)P(X = t_i)}{f_Y(u)} \qquad \text{for} \quad f_Y(u) > 0.$$

7. Proofs of the properties of conditional expectation

In this section, we show that if $e(\cdot)$ is a Borel function which satisfies the defining relation A) $E[I_M(Y)e(Y)] = E[I_M(Y)g(X)]$ for any Borel set M on the codomain of Y, then properties CE2) through CE13) hold for $e(Y) = E[g(X)|Y]$. Note that when we write $e(Y) = E[g(X)|Y]$, we are asserting that $e(\cdot)$ satisfies the defining relation (A) and must therefore be unique a.s. $[P_Y]$.

In the proofs, we employ the properties E1) through E6) of mathematical expectation. Actually, we need only the simpler part a) of property E6). Note that the proofs do not involve the complexities of conditional distributions. The reader who wishes to go through the proofs carefully may wish to use the summary of properties of mathematical expectation in Appendix I. A tally of the use of these properties might be instructive. To simplify writing, we drop the "a.s." in many places.

At several places, the arguments require an acquaintance with measure-theoretic ideas beyond that assumed of most readers. In these instances, we sketch the ideas of the proofs, in order to indicate to the interested reader what to look for in seeking a more complete treatment. The goal is insight into the mathematical structure as an aid to interpretation and application.

CE2) Linearity.

Let $e_1(Z) = E[g(X)|Z]$, $e_2(Z) = E[h(Y)|Z]$, $e(Z) = E[ag(X) + bh(Y)|Z]$.

For any Borel set M in the codomain of Z, we have

$$E\{I_M(Z)[ag(X) + bh(Y)]\} = E[I_M(Z)e(Z)] \qquad \text{by CE1).}$$

Also

$$\begin{aligned} E\{I_M(Z)[ag(X) + bh(Y)]\} &= aE[I_M(Z)g(X)] + bE[I_M(Z)h(Y)] && \text{by E2)} \\ &= aE[I_M(Z)e_1(Z)] + bE[I_M(Z)e_2(Z)] && \text{by CE1)} \\ &= E\{I_M(Z)[ae_1(Z) + be_2(Z)]\} && \text{by E2).} \end{aligned}$$

By E6), we have $e(Z) = ae_1(Z) + be_2(Z)$ a.s. []

CE3) Positivity; monotonicity.

$g(X) \geq 0$ a.s. implies $E[I_M(Y)g(X)] \geq 0$ by E3)

implies $E[I_M(Y)e(Y)] \geq 0$ by CE1).

Suppose $e(Y) < 0$ for $\omega \in A$. Then there is a Borel set M_0 with $A = Y^{-1}(M_0)$. Thus, $I_{M_0}(Y)e(Y) = I_A e(Y) \leq 0$. By E3), we have $E[I_{M_0}(Y)e(Y)] \leq 0$, with equality iff $I_A e(Y) = 0$ a.s.. But this requires $P(A) = 0$, which is equivalent to the condition $e(Y) \geq 0$ a.s.

Monotonicity follows from positivity and linearity. []

CE4) Monotone convergence.

Consider the nondecreasing case $X_n \uparrow X$ a.s. Put $e_n(Y) = E[X_n|Y]$ and $e(Y) = E[X|Y]$. Then by CE3), $e_n(Y) \leq e_{n+1}(Y) \leq e(Y)$ a.s., all $n \geq 1$. The almost-sure restriction means that we can neglect an event (set of ω) of zero probability and have the indicated relationship for all other ω. By ordinary rules of limits, for any ω other than the exceptional set, we we have $e^*(Y) = \lim_n e_n(Y) \leq e(Y)$, which means the inequalities hold a.s. For any Borel set M, $I_M(Y)X_n \uparrow I_M(Y)X$ a.s., and $I_M(Y)e_n(Y) \uparrow I_M(Y)e^*(Y)$ a.s. so that by monotone convergence for expectation,

$E[I_M(Y)e_n(Y)] = E[I_M(Y)X_n] \uparrow E[I_M(Y)X] = E[I_M(Y)e(Y)]$ and

$E[I_M(Y)e_n(Y)] \uparrow E[I_M(Y)e^*(Y)]$. Hence,

$E[I_M(Y)e^*(Y)] = E[I_M(Y)e(Y)]$ for all Borel sets M on the codomain of Y. This ensures $e^*(Y) = e(Y)$ a.s., by E6). []

CE5) Independence. a) $\{X,Y\}$ is independent iff b) $E[I_N(X)|Y] = E[I_N(X)]$ a.s. for all Borel N iff c) $E[g(X)|Y] = E[g(X)]$ a.s. for all Borel functions g.

a) $\Rightarrow$ c) $\{g(X), I_M(Y)\}$ is independent; hence

$$E[I_M(Y)g(X)] = E[I_M(Y)]E[g(X)] \quad \text{by E5)}$$

$$= E\{E[g(X)]I_M(Y)\} \quad (E[g(X)] \text{ a constant}) \quad \text{by E2)}$$

$$E[I_M(Y)g(X)] = E[I_M(Y)e(Y)] \quad \text{by CE1).}$$

Since the constant $E[g(X)]$ is a Borel function of Y, we conclude by E6) that $e(Y) = E[g(X)]$ a.s.

c) $\Rightarrow$ b), since b) is a special case of c)

b) $\Rightarrow$ a) For any Borel sets M, N on the codomains of X, Y, respectively,

$$E[I_M(X)I_N(Y)] = E\{I_N(Y)E[I_M(X)]\} \quad \text{by hyp. and CE1)}$$

$$= E[I_M(X)]E[I_N(Y)] \quad \text{by E2)}$$

which ensures independence of $\{X,Y\}$ by E5). []

CE6) Extension of CE1) to general Borel functions.

First we suppose $g \geq 0$. By positivity CE3), we have $e(Y) \geq 0$ a.s.

1) By CE1), the proposition is true for $h = I_M$.

2) By linearity CE2), the proposition is true for any simple function $h = \sum_{i=1}^{m} t_i I_{M_i}$.

3) For $h \geq 0$, there is a sequence of simple functions $h_n \uparrow h$. This implies $h_n(Y)g(X) \uparrow h(Y)g(X)$ and $h_n(Y)e(Y) \uparrow h(Y)e(Y)$ a.s. Hence, by monotone convergence E4), for expectations,

$$E[h_n(Y)g(X)] \uparrow E[h(Y)g(X)] \text{ and } E[h_n(Y)e(Y)] \uparrow E[h(Y)e(Y)].$$

Since for each n, $E[h_n(Y)g(X)] = E[h_n(Y)e(Y)]$, the limits must be the same.

4) For general Borel h, we have $h = h_+ - h_-$, where both h_+ and h_- are nonnegative Borel functions. By linearity and 3), we have

$$E[h(Y)g(X)] = E[h_+(Y)g(X)] - E[h_-(Y)g(X)]$$

$$= E[h_+(Y)e(Y)] - E[h_-(Y)e(Y)] = E[h(Y)e(Y)].$$

5) For general Borel g, we have $g = g_+ - g_-$, where both g_+ and g_- are nonnegative Borel functions. By linearity and 4), we have

$$E[h(Y)g(X)] = E[h(Y)g_+(X)] - E[h(Y)g_-(X)]$$
$$= E[h(Y)e_+(Y)] - E[h(Y)e_-(Y)] = E[h(Y)e(Y)],$$

where $e_+(Y) = E[g_+(X)|Y]$, $e_-(Y) = E[g_-(X)|Y]$, and

$e(Y) = e_+(Y) - e_-(Y)$ a.s., by CE2). []

CE7) If $X = h(Y)$, then $E[g(X)|Y] = g(X)$ a.s.

$g(X) = g[h(Y)] = h^*(Y)$, with h^* Borel. For any Borel set M,

$E[I_M(Y)g(X)] = E[I_M(Y)h^*(Y)] = E[I_M(Y)e(Y)]$ by CE1).

But this ensures

$h^*(Y) = e(Y)$ a.s. by E6). []

CE8) $E[h(Y)g(X)|Y] = h(Y)E[g(X)|Y]$ a.s.

For any Borel set M, $I_M(Y)h(Y)$ is a Borel function of Y. Set

$e(Y) = E[g(X)|Y]$ and $e^*(Y) = E[h(Y)g(X)|Y]$.

Now $E[I_M(Y)h(Y)g(X)] = E[I_M(Y)h(Y)e(Y)]$ by CE6)

and $E[I_M(Y)h(Y)g(X)] = E[I_M(Y)e^*(Y)]$ by CE1).

Hence, $h(Y)e(Y) = e^*(Y)$ by E6). []

CE9) If $Y = h(W)$, then $E\{E[g(X)|Y]|W\} = E\{E[g(X)|W]|Y\} = E[g(X)|Y]$ a.s.

Set $e(Y) = E[g(X)|Y] = e[h(W)] = h^*(W)$ and $e^*(W) = E[g(X)|W]$.

Then, $E\{E[g(X)|Y]|W\} = E[h^*(W)|W] = h^*(W) = e(Y)$ by CE7).

For any Borel set M on the codomain of Y, let $N = h^{-1}(M)$. By Theorem A1-1, $I_M(Y) = I_N(W)$. Repeated use of CE1) gives

$$E[I_M(Y)g(X)] = E[I_M(Y)e(Y)]$$
$$= E[I_N(W)g(X)] = E[I_N(W)e^*(W)] = E[I_M(Y)e^*(W)]$$
$$= E\{I_M(Y)E[e^*(W)|Y]\}.$$

Hence $e(Y) = E[e^*(W)|Y]$ a.s. by E6). []

Proof of CE10) requires some results of measure theory beyond the scope of the present work. We establish the proposition first for the special case that X, Y are real-valued, with joint density function f_{XY}; then we sketch the ideas of a general proof.

CE10) $E[g(X,Y)|Y = u] = E[g(X,u)|Y = u]$ a.s. $[P_Y]$.

PROOF FOR SPECIAL CASE. X,Y have joint density f_{XY}.

Let $f_{X|Y}$ be defined as in Sec C2. Put $e(u,v) = E[g(X,v)|Y = u]$ and $e^*(u) = E[g(X,Y)|Y = u]$. Then $E[I_M(Y)g(X,Y)] = E[I_M(Y)e^*(Y)]$ $\forall$ Borel set M. This is equivalent to

$$\iint I_M(u)g(t,u)f_{XY}(t,u)\,dtdu = \int I_M(u)e^*(u)\,du.$$

The left-hand integral may be written

$$\int I_M(u)\left[\int g(t,u)f_{X|Y}(t|u)\,dt\right]f_Y(u)\,du = \int I_M(u)e(u,u)f_Y(u)\,du.$$

Thus,

$$\int I_M(u)e(u,u)f_Y(u)\,du = \int I_M(u)e^*(u)f_Y(u)\,du \quad \text{or, equivalently,}$$

$$E[I_M(Y)e(Y,Y)] = E[I_M(Y)e^*(Y)] \quad \forall \text{ Borel set } M.$$

We conclude $e(Y,Y) = e^*(Y)$ a.s. by E6).

IDEA OF A GENERAL PROOF

If the theorem can be established for $g(t,u) = I_Q(t,u)$, where Q is any Borel set on the codomain of (X,Y), then a "standard argument" such as used in the proof of CE6) extends the theorem to any Borel function g such that $E[g(X,Y)]$ is finite.

We first consider Borel sets of the form $Q = M \times N$, where M, N are Borel sets in the codomains of X, Y, respectively. Then $I_Q(t,u) = I_M(t)I_N(u)$.

Let $e(u,v) = E[g(X,v)|Y = u] = E[I_M(X)I_N(v)|Y = u] = I_N(v)E[I_M(X)|Y = u]$ and $e^*(u) = E[g(X,Y)|Y = u] = E[I_M(X)I_N(Y)|Y = u]$.

Now $e(Y,Y) = I_N(Y)E[I_M(X)|Y]$ and $e^*(Y) = I_N(Y)E[I_M(X)|Y]$ a.s. by CE8).

Hence, $e(Y,Y) = e^*(Y)$ a.s. or $e(u,u) = e^*(u)$ a.s. $[P_Y]$.

By linearity, the equality holds for any Borel set $Q = \biguplus_{i=1}^{n} M_i \times N_i$, since in this case $I_Q = \sum_{i=1}^{n} I_{M_i} I_{N_i}$. Hence, equality holds for the class $\mathcal{B}_0$ consisting of all finite, disjoint unions of sets of the form $M \times N$. A number of arguments may be used to show that equality holds for all Borel sets Q. We sketch one proof. It is known that the class $\mathcal{B}_0$ is a field and that the minimal sigma field which contains it is the class $\mathcal{B}$ of Borel sets. Let $\mathcal{B}^*$ be the class of sets for which equality holds. If $\{Q_i: 1 \leq i\}$ is a monotone class of sets in $\mathcal{B}^*$, the sequence $\{I_{Q_i}: 1 \leq i\}$ is a monotone sequence of Borel functions. Use of monotone convergence E4) for expectations shows that equality holds for I_Q where Q is the limit of the sequence Q_i. Thus, $\mathcal{B}^*$ is a monotone class. By a well known theorem, $\mathcal{B}^*$ must include $\mathcal{B}$. This means that equality holds for every Borel set Q. []

CE11) If $\{X,Y\}$ is independent, $E[g(X,Y)|Y = u] = E[g(X,u)]$ a.s. $[P_Y]$

By CE5), independence of $\{X,Y\}$ ensures $e(u,v) = E[g(X,v)|Y = u] = E[g(X,v)]$ a.s. $[P_Y]$, so $e^*(u) = e(u,u) = E[g(X,u)]$ a.s. $[P_Y]$. []

CE12) Triangle inequality.

Since $g(X) \leq |g(X)|$, we have $E[g(X)|Y] \leq E[|g(X)||Y]$ a.s. by CE3).

Since $-g(X) \leq |g(X)|$, we have $-E[g(X)|Y] \leq E[|g(X)||Y]$ a.s. by CE3), CE2).

Hence, $|E[g(X)|Y]| \leq E[|g(X)||Y]$ a.s. []

CE13) Jensen's inequality.

Convex function g satisfies $g(t) \geq g(y) + \lambda(y)(t - y)$, where λ is a nondecreasing function. Set $e(Y) = E[X|Y]$. Then

$g(X) \geq g[e(Y)] + \lambda[e(Y)][X - e(Y)]$. If we take conditional expectation,

$$
\begin{aligned}
E[g(X)|Y] &\geq E\{g[e(Y)] + \lambda[e(Y)][X - e(Y)]|Y\} \quad \text{a.s.} && \text{by CE3)} \\
&= E\{g[e(Y)]|Y\} + E\{\lambda[e(Y)]X|Y\} - E\{\lambda[e(Y)]e(Y)|Y\} && \text{by CE2)} \\
&= g[e(Y)] + \lambda[e(Y)]e(Y) - \lambda[e(Y)]e(Y) && \text{by CE7), CE8)} \\
&= g\{E[X|Y]\} \quad \text{a.s.}
\end{aligned}
$$

[]

8. Problems

C-1 Prove parts b) and c) of Theorem C1-1.

C-2 Suppose X, Y have joint density function f_{XY}. Use Theorem C1-2 and property E1a) for expectation to show that

a) If $P(Y \in M) > 0$, then

$$E[g(X)|Y \in M] = \int_M [\int g(t) f_{XY}(t,u)\, dt]\, du \;/\; \int_M [\int f_{XY}(t,u)\, dt]\, du.$$

b) If $P(X \in N) > 0$, then

$$E[g(X)|X \in N] = \int_N g(t) f_X(t)\, dt \;/\; \int_N f_X(t)\, dt$$

C-3 Show $E[g(X)|A] = E[g(X)|AB]P(B|A) + E[g(X)|AB^c]P(B^c|A)$.

C-4 If X is discrete and Y is absolutely continuous, then the joint distribution can be described by a hybrid mass-density function f_{XY}, such that $P(X = t_i, Y \in M) = \int_M f_{XY}(t_i,u)\, du$. Develop an expression for $e(u) = E[g(X)|Y = u]$ in this case.

C-5 Let $X = 0I_{A_1} + 2I_{A_2} + 3I_{A_3}$ and $Y = I_{B_1} + 3I_{B_2}$ (canonical form), with joint probability distribution such that $P(A_1B_1) = 1/6$, $P(A_2B_2) = 1/2$, and $P(A_3B_1) = 1/3$. Show that $E[X|Y = 1] = E[X|Y = 3] = E[X]$, but that $\{X,Y\}$ is not independent.

C-6 Show that for X real, the triangle inequality is a special case of Jensen's inequality.

C-7 Suppose $\{X,Y\}$ is independent, and each random variable is uniform on $[-1, 1]$. Let $Z = g(X,Y)$ be given by

$$Z = \begin{cases} X & \text{for } X^2 + Y^2 \leq 1 \\ c & \text{for } X^2 + Y^2 > 1. \end{cases}$$

Determine $E[Z|X^2 + Y^2 \leq 1]$ and $E[Z|X^2 + Y^2 > 1]$. Use these results to determine $E[Z]$.

C-8 X,Y have joint density function $f_{XY}(t,u) = \frac{8}{9} tu$ for $1 \leq t \leq u \leq 2$ (and zero elsewhere). Determine

a) $E[X^2 + Y^2|X = t]$ b) $E[XY|X = t]$ c) $E[X|X \leq \frac{1}{2}(Y + 1)]$.

C-9 The pair $\{X,Y\}$ is independent, with $f_X(t) = f_Y(t) = 1/2$ for $-1 \leq t \leq 0$, $= \frac{1}{2} e^{-t}$ for $0 \leq t$ (and zero elsewhere). Determine

a) $E[X^2 + Y^2 | X = t]$ b) $E[XY | X = t]$.

C-10 Use the fact that $g(X,Y) = g^*(X,Y,Z)$, with g^* Borel if g is, to establish the following extension of CE10).

$$E[g(X,Y) | Y = u, Z = v] = E[g(X,u) | Y = u, Z = v] \quad \text{a.s. } [P_{YZ}].$$

C-11 Use CE9a) and the result of problem C-10 to show that if $F_{Y|Z}$ is a regular conditional distribution function, then

$$E[g(X,Y) | Z = v] = \int E[g(X,u) | Y = u, Z = v] \, dF_{Y|Z}(u|v) \quad \text{a.s. } [P_Z].$$

C-12 Suppose X is a real random variable with $E[X^2]$ finite. Let $e(Y) = E[X|Y]$ and $v(Y) = \text{Var}[X|Y] = E\{[X - e(Y)]^2 | Y\}$. Show that

a) $v(Y) = E[X^2|Y] - E^2[X|Y] = E[X^2|Y] - e^2(Y)$

b) $\text{Var}[e(Y)] = E[e^2(Y)] - E^2[X] = E(E^2\{X - E[X]|Y\})$

c) $\text{Var}[X] = E[v(Y)] + \text{Var}[e(Y)] = E\{\text{Var}[X|Y]\} + \text{Var}\{E[X|Y]\}$.

C-13 The following is a model for the demand of a random number of "customers", who buy independently but with the same individual demand probabilities. Suppose

i) $\{X_k : 1 \leq k\}$ is iid (independent, identically distributed), with $E[X_k^2]$ finite (individual demands).

ii) N is a nonnegative, integer-valued random variable with $E[N^2]$ finite (number of customers).

iii) $\{N, X_k : 1 \leq k\}$ is an independent class.

$$D = \begin{cases} 0 & \text{for } N = 0 \\ \sum_{k=1}^{n} X_k = Y_n & \text{for } N = n > 0 \end{cases} \quad \text{(composite demand).}$$

If $A_n = \{\omega : N(\omega) = n\}$, then $D = \sum_{n=0}^{\infty} I_{A_n} Y_n = \sum_{n=0}^{\infty} I_{\{n\}}(N) Y_n$.

a) Show $E[D|N = n] = nE[X]$ and $\text{Var}[D|N = n] = n\text{Var}[X]$

Note. $E[D|N = n] = E[I_{\{n\}}(N)D]/P(N = n)$, etc.

b) Show $E[D] = E[N]E[X]$.

c) Use the result of problem C-12 to show

$$Var[D] = E[N]Var[X] + Var[N]E^2[X].$$

d) Suppose N is Poisson (λ) and X is uniform on $[0,a]$. Calculate $E[D]$ and $Var[D]$.

C-14 The characteristic function φ_X for a real random variable X is $\varphi_X(u) = E[e^{iuX}]$, defined for all real u (i is the complex imaginary unit, $i^2 = -1$). The generating function g_N for a nonnegative, integer-valued random variable N is $g_N(s) = E[s^N] = \sum_{k=0}^{\infty} s^k P(N = k)$, defined at least for $|s| < 1$, although possibly for a much larger domain. It is readily shown that addition of a finite number of members of an independent class of random variables corresponds to multiplying their characteristic functions (or their generating functions, if they exist). Consider the composite random variable D in problem C-13.

a) Show that $\varphi_D(u) = g_N[\varphi_X(u)]$, where g_N is the generating function for N and φ_X is the common characteristic function for the X_k. [Suggestion. Condition by N, then take expectation. $E[e^{iuD}|N = n] = E[e^{iuY_n}]$].

b) Show that if the X_k are nonnegative, integer-valued with common generating function g_X, then $g_D(s) = g_N[g_X(s)]$.

c) Suppose N is Poisson (λ). Show that $g_N(s) = \exp[\lambda(s - 1)]$, so that $\varphi_D(u) = \exp\{\lambda[\varphi_X(u) - 1]\}$.
[$P(N = k) = e^{-\lambda}\frac{\lambda^k}{k!}$].

C-15 The correlation ratio of X with respect to Y (see Rényi [1970], p 275 ff) is $K[X|Y] = \sigma[e(Y)]/\sigma[X]$ (see Problem C-12). Show that the following properties hold:

a) $0 \leq K[X|Y] \leq 1$

b) If $\{X,Y\}$ is independent, then $K[X|Y] = 0$.

c) $K[X|Y] = 1$ iff there is a Borel function g with $X = g(Y)$.

d) $K^2[X|Y] = \sup_g \rho^2[X, g(Y)]$, where g ranges over the set of Borel functions such that $E[g^2(Y)]$ is finite and ρ is the correlation coefficient for X and $g(Y)$.

e) $K^2[X|Y] = \rho^2[X, g(Y)]$ iff there exist a, b $(a \neq 0)$ such that $g(Y) = ae(Y) + b$ a.s.

Suggestion. For d), e), use CE1b) and Schwarz' inequality. Work with standardized random variables obtained by subtracting the mean and dividing by the standard deviation.

C-16 Suppose f and g are Borel functions such that $E[f(X)] = E[g(Y)] = 0$, $Var[f(X)] = Var[g(Y)] = 1$, and $E[f(X)g(Y)] = \sup_{\varphi,\psi} E[\varphi(X)\psi(Y)] = \lambda$. Show that

a) $E[f(X)|Y] = \lambda g(Y)$ a.s. and $E[g(Y)|X] = \lambda f(X)$ a.s.

b) $E\{E[f(X)|Y]|X\} = \lambda^2 f(X)$ a.s. and $E\{E[g(Y)|X]|Y\} = \lambda^2 g(Y)$ a.s.

c) $E[f(X)|g(Y)] = \lambda g(Y)$ a.s. and $E[g(Y)|f(X)] = \lambda f(X)$ a.s.

Suggestion. Use CE1b) and Schwarz' inequality.

D. Conditional Independence, Given a Random Vector

D. CONDITIONAL INDEPENDENCE, GIVEN A RANDOM VECTOR

D. Conditional independence, given a random vector

The concept of conditional independence of random vectors which we consider in the following sections has been utilized widely in advanced treatments of Markov processes. In such treatments, the concept is usually expressed in terms of conditional expectation, given a sigma field of events. Our immediate goal is to reformulate the essential features of such treatments in terms of the more readily accessible conditional expectation, given a random vector, developed in the preceding sections. We then illustrate the usefulness of the conditional independence concept by showing how it appears naturally in certain problems in decision theory. In Sec E1 and following, we apply the concept to Markov processes.

1. The concept and some basic propositions

Although historically there seems to be no connection, it may be instructive to consider how the concept of conditional independence, given a random vector, may be seen as an extension of the simpler concept of conditional independence, given an event. Suppose $\{A,B\}$ is conditionally independent, given C, with $A = X^{-1}(M)$ and $B = Y^{-1}(N)$. Then the product rule $P(AB|C) = P(A|C)P(B|C)$ may be expressed $E[I_M(X)I_N(Y)|C] = E[I_N(X)|C]E[I_N(Y)|C]$ If this rule holds for all Borel sets M, N on the codomains of X, Y, respectively, we should be inclined to say $\{X,Y\}$ is conditionally independent, given C. Suppose Z is a simple random variable, with $C = \{Z = z_k\}$. Then, in these terms, we would say $\{X,Y\}$ is conditionally independent, given $Z = z_k$. If the product rule holds for all Borel sets M, N in the codomains of X, Y, respectively, and for all z_k in the range of Z, we should then be inclined to say the pair $\{X,Y\}$ is conditionally independent, given Z. With the aid of the result of example C3-a, we may give this set of conditions a simple formulation which points to a general

definition. We have

$$E[I_M(X)I_N(Y)|Z] = \sum_k E[I_M(X)I_N(Y)|Z = z_k]I_{C_k}$$

with similar expressions for $E[I_M(X)|Z]$ and $E[I_N(Y)|Z]$. Using the facts that $I_{C_j}I_{C_k} = 0$ for $j \neq k$ and $I_{C_k}^2 = I_{C_k}$, we obtain

$$E[I_M(X)|Z]E[I_N(Y)|Z] = \sum_k E[I_M(X)|Z = z_k]E[I_N(Y)|Z = z_k]I_{C_k}.$$

We thus have

i) $E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z]E[I_N(Y)|Z]$ iff

ii) $E[I_M(X)I_N(Y)|Z = z_k] = E[I_M(X)|Z = z_k]E[I_N(Y)|Z = z_k]$ for all k.

We have seen above that the set of conditions ii) is a reasonable basis for the notion that $\{X,Y\}$ is conditionally independent, given simple random vector Z. This development suggests the simpler equivalent expression i) may be the more useful way to characterize the condition. Further evidence is provided by the following set of equivalent conditions (see Sec D5 for proofs).

For any random vector Z, the following conditions are equivalent:

CI1) $E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z]E[I_N(Y)|Z]$ a.s. $\forall$ Borel sets M, N

CI2) $E[I_M(X)|Z,Y] = E[I_M(X)|Z]$ a.s. $\forall$ Borel sets M

CI3) $E[I_M(X)I_Q(Z)|Z,Y] = E[I_M(X)I_Q(Z)|Z]$ a.s. $\forall$ Borel sets M, Q

CI4) $E[I_M(X)I_Q(Z)|Y] = E\{E[I_M(X)I_Q(Z)|Z]|Y\}$ a.s. $\forall$ Borel sets M, Q

CI5) $E[g(X)h(Y)|Z] = E[g(X)|Z]E[h(Y)|Z]$ a.s. $\forall$ Borel functions g, h

CI6) $E[g(X)|Z,Y] = E[g(X)|Z]$ a.s. $\forall$ Borel functions g

CI7) $E[g(X,Z)|Z,Y] = E[g(X,Z)|Z]$ a.s. $\forall$ Borel functions g

CI8) $E[g(X,Z)|Y] = E\{E[g(X,Z)|Z]|Y\}$ a.s. $\forall$ Borel functions g

Several facts should be noted. For one thing, properties CI5) through CI8) are generalizations of CI1) through CI4), respectively, in that the indicator functions are replaced by real-valued Borel functions, subject only to the restriction that the resultant random variables $g(X)$, $h(Y)$, and $g(X,Y)$ should have finite expectations. It is desirable to have the properties CI1) through CI4) included in the list of equivalences to show that it is sufficient to establish one of these simpler conditions in order to be able to assert the apparently more general counterparts CI5) through CI8), respectively.

Expressions CI2) and CI6) show that if X is conditioned by Z, further conditioning by Y has no appreciable effect. We thus have an analogy to the idea that the event pair $\{A,B\}$ is conditionally independent, given event C, if once A is conditioned by C, then further conditioning by B has no effect on the likelihood of the occurrence of A. Expressions CI3) and CI7) generalize this to say that if (X,Z) is conditioned by Z, further conditioning by Y has no appreciable effect. It is clear that the role of X and Y could be interchanged in these statements. Conditions CI4) and CI8) have no counterpart in the theory of conditional independence of events. They do, however, play an important role in the theory of the new concept; and they include as a special case the Chapman-Kolmogorov equation which plays a prominent role in the theory of Markov processes (cf Sec E4).

These considerations indicate that we have identified a potentially useful concept which is properly named conditional independence. We can use any of the eight equivalent propositions as the basis for definition. As in the case of independence of events and of random variables, we use the product rule CI1).

DEFINITION. The pair of random vectors $\{X,Y\}$ is <u>conditionally independent, given</u> Z, iff the product rule CI1) holds.

An arbitrary class of random vectors is conditionally independent, given Z, iff an analogous product rule holds for each finite subclass of two or more members of the class.

If the pair $\{X,Y\}$ is conditionally independent, given Z, we should expect that any Borel functions of these two variables should be conditionally independent. This is the case.

CI9) If $\{X,Y\}$ is conditionally independent, given Z, $U = h(X)$, and $V = k(Y)$, with h, k Borel, then $\{U,V\}$ is conditionally independent, given Z.

For convenience of reference, we list several additional properties of conditional independence utilized in various subsequent developments.

CI10) If the pair $\{X,Y\}$ is conditionally independent, given Z, then

a) $E[g(X)h(Y)] = E\{E[g(X)|Z]E[h(Y)|Z]\} = E[e_1(Z)e_2(Z)]$, and

b) $E[g(X)|Y \in N]P(Y \in N) = E\{E[I_N(Y)|Z]E[g(X)|Z]\}$

CI11) If $\{Y, (X,Z)\}$ is independent, then $\{X,Y\}$ is conditionally independent, given Z.

CI12) If $\{X,Y\}$ is conditionally independent, given Z, then

$E[g(X,Y)|Y = u, Z = v] = E[g(X,u)|Z = v]$ a.s. $[P_{YZ}]$

Proofs of these propositions are provided in Sec D5.

2. Some elements of Bayesian analysis

Classical statistics postulates a population distribution to be determined by sampling, or some other appropriate form of experimentation. Typically, the distribution is supposed to belong to a specified class (e.g., normal, exponential, binomial, Poisson, etc.) which is characterized by certain parameters. A finite set of parameters can be viewed as a set of coordinates for a single vector-valued parameter. The value θ of the parameter is assumed fixed, but is unknown. Hence, there is uncertainty about its value.

An alternative formulation results from modeling the uncertainty in a probabilistic manner. The uncertain value of the parameter is viewed as the value of a random vector; i.e., $\theta = H(\omega)$. The value $H(\omega)$ of the parameter random vector H reflects the state of nature. If X is a random variable representing the population, then the distribution for X is determined by the value $\theta = H(\omega)$ of the parameter random vector. To carry out statistical analysis, we must characterize appropriately the joint distribution for the pair $\{X,H\}$. This is usually done by assuming a conditional distribution for X, given H, represented by conditional distribution function $F_{X|H}$ (or an appropriate alternative); and by utilizing any information about the probable values of the parameter to determine a prior distribution for H, represented by a distribution function F_H (or some appropriate alternative).

A central notion of classical statistics is a random sample of size n. Some sampling act, or survey, or experiment is done repeatedly, in such a way that the outcome of one sampling act does not affect operationally the outcome of any other. This is modeled as a class $\{X_1, X_2, \ldots, X_n\}$ of independent random variables, each having the population distribution.

Each sampling act corresponds to the observation of one of the random variables in the sample. A random sample is a finite case of an arbitrary iid (independent, identically distributed) class $\{X_i : i \in J\}$.

Under the new point of view, the appropriate assumption seems to be that the class $\{X_i : i \in J\}$ is conditionally independent, given H, with all random variables X_i having the same conditional distribution, given H. Under a given state of nature, the result of taking an observation of X_i does not affect and is not affected by the results of observing any combination of the other random variables. We find it convenient to adopt the following terminology:

DEFINITION. A class $\{X_i : i \in J\}$ is ciid, given H, iff the class is conditionally independent, and each random variable X_i has the same conditional distribution, given H. A random sample (of size n), given H, is a finite class $\{X_i : 1 \leq i \leq n\}$ which is ciid, given H.

Let us see what this means for the conditional distribution functions. To simplify writing, put $W = (X_1, X_2, \ldots, X_n)$ and let $I_t = I_{N_t}$, where $N_t = (-\infty, t]$. Then

$$\begin{aligned} F_{W|H}(t_1, t_2, \ldots, t_n|u) &= P(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_n \leq t_n|H = u) \\ &= E[I_{t_i}(X_1) I_{t_2}(X_2) \cdots I_{t_n}(X_n)|H = u] \\ &= \prod_{i=1}^{n} E[I_{t_i}(X_i)|H = u] \quad \text{by conditional independence} \\ &= \prod_{i=1}^{n} F_{X|H}(t_i|u). \end{aligned}$$

Thus, the conditional distribution function obeys the product rule. Partial differentiation by the t_i shows that the conditional density, when it exists, also satisfies the product rule

$$f_{W|H}(t_1, t_2, \ldots, t_n|u) = \prod_{i=i}^{n} f_{X|H}(t_i|u).$$

Bernoulli trials, given H. We illustrate these ideas by considering the important special case of Bernoulli trials. A sequence of "identical" trials is performed in an operationally independent manner. Let E_i = event of a "success" on the ith trial in the sequence, and set $X_i = I_{E_i}$, so that X_i has the property that it takes on the value 1 if E_i occurs and takes on the value 0 if E_i fails to occur (E_i^c occurs). On a given sequence of trials, the probability p of success on a trial does not vary with i. Now p is a parameter, representing a state of nature. We model it as the value of a parameter random variable H with the interval [0,1] as its range. For a given value of H, the results of the various trials are conditionally independent. Thus, we assume $\{X_i : 1 \leq i\}$ is ciid, given H. Let $I_{\{0\}}$ be the indicator function for the set $\{0\}$ and similarly for $I_{\{1\}}$. We suppose

$$P(E_i|H = u) = P(X_i = 1|H = u) = E[I_{\{1\}}(X_i)|H = u] = u \qquad 0 \leq u \leq 1$$

$$P(E_i^c|H = u) = P(X_i = 0|H = u) = E[I_{\{0\}}(X_i)|H = u] = 1 - u.$$

These assumptions ensure

$$E[X_i|H = u] = e(u) = u \quad \text{a.s. } [P_H].$$

It is convenient in this case to say the sequence is Bernoulli, given H = u. To see how analysis of such sequences relates to analysis of ordinary Bernoulli sequences, suppose, for example, we observe the sequence $E_1E_2^cE_3^c$. Then

$$\begin{aligned} P(E_1E_2^cE_3^c|H = u) &= E[I_{\{1\}}(X_1)I_{\{0\}}(X_2)I_{\{0\}}(X_3)|H = u] \\ &= E[I_{\{1\}}(X_1)|H = u]E[I_{\{0\}}(X_2)|H = u]E[I_{\{0\}}(X_3)|H = u] \\ &= u(1 - u)(1 - u). \end{aligned}$$

The product after the second equality sign is a result of conditional independence. The pattern here is obviously the same as in the analysis of ordinary Bernoulli trials, except that we write u for p. To obtain the conditional probability of any such sequence, given H = u, include a

factor u for each uncomplemented E_i and a factor $1 - u$ for each complemented E_i.

The random variable $S_n = X_1 + X_2 + \ldots + X_n$ counts the number of "successes" in the first n trials of a sequence. In the ordinary case, S_n has the binomial distribution with parameters (p, n). As the discussion above indicates, if the sequence is Bernoulli, given $H = u$, we simply replace p by u in the analysis of the ordinary case to obtain

$$P(S_n = k \mid H = u) = C(n,k)u^k(1 - u)^{n-k} \qquad 0 \leq u \leq 1.$$

We still have the problem of determining the distribution of H (i.e., the prior distribution with which to begin analysis). Partly because of a well-known integral formula

$$\int_0^1 u^r(1 - u)^s \, du = \frac{r!\, s!}{(r + s + 1)!} = 1/(r + s + 1)C(r+s,r).$$

A commonly employed class of distributions is the class of

Beta distributions

Real random variable H has the Beta distribution with parameters $(a+1, b+1)$ (a, b nonnegative integers) iff it has the density function

$$f_H(t) = \begin{cases} \dfrac{(a + b + 1)!}{a!\, b!}\, t^a(1 - t)^b & 0 \leq t \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $a = b = 0$, the distribution is uniform on $[0,1]$.

Straightforward calculations show that f_H has a maximum at $t = a/(a + b)$, an

$$E[H] = \frac{a + 1}{a + b + 2} \qquad \mathrm{Var}[H] = \frac{(a + 1)(b + 1)}{(a + b + 2)^2(a + b + 3)}$$

$$E[H^2] = \frac{(a + 1)(a + 2)}{(a + b + 2)(a + b + 3)}$$

$$E[H^k] = \frac{(a + 1)(a + 2) \ldots (a + k)}{(a + b + 2)(a + b + 3) \ldots (a + b + k + 1)}.$$

If prior knowledge indicates that the value of H lies in a certain part of the unit interval, with a degree of certainty reflected in the size of

the variance, the parameters a, b may be adjusted to reflect these conditions. If there is no prior knowledge favoring any set of probable values, the complete ignorance may be expressed by taking the uniform case (a = b = 0)

Example D2-a

A quantity of n items from one run of a production line is selected at random for testing. There is probability p that any device in the lot will meet specifications. The quantity p is constant over any one run; its value depends on how well the manufacturing process, including selection or preparation of raw materials is controlled. We wish to estimate the parameter p from tests and prior knowledge.

SOLUTION.

We adopt the point of view that p is the value of a parameter random variable H. Past experience indicates a reasonable prior distribution is Beta, with parameters (3,2) -- i.e., a = 2, b = 1. Thus $f_H(t) = 12\, t^2(1 - t)$, $0 \leq t \leq 1$ (maximum at $t = 2/3$). Then

$$P(E_i) = E[X_i] = E\{E[X_i|H]\} = E[e(H)] = E[H] = \frac{a+1}{a+b+2} = 3/5$$

Suppose $X_1 = 1$. Then

$$P(E_2|X_1 = 1) = P(E_1E_2)/P(E_1) = E[X_1X_2]/E[X_1] = E[e(H)e(H)]/E[X_1] \quad \text{by CI9}$$

$$= E[H^2]/E[H] = \frac{(a+1)(a+2)}{(a+b+2)(a+b+3)} \frac{(a+b+2)}{(a+1)}$$

$$= \frac{a+2}{a+b+3} = 4/6 = 2/3.$$

Note that $\{E_1, E_2\}$ is not an independent pair, since $P(E_2|E_1) \neq P(E_2)$. []

Suppose a prior distribution for H is assumed. A sequence of n trials is performed. It is desired to update the distribution for H on the basis of the results of this experiment. Suppose k successes occur (i.e., $S_n = k$); we want to determine the conditional distribution for H, given $S_n = k$. Now

$$F_{H|S_n}(u|k) = P(H \le u|S_n = k) = E[I_{N_u}(H)I_{\{k\}}(S_n)]/E[I_{\{k\}}(S_n)]$$

$$= E\{I_{N_u}(H)E[I_{\{k\}}(S_n)|H]\}/E\{E[I_{\{k\}}(S_n)|H]\} \text{ , where}$$

$$E[I_{\{k\}}(S_n)|H = u] = P(S_n = k|H = u) = C(n,k)u^k(1 - u)^{n-k} \quad 0 \le u \le 1.$$

Suppose H has the Beta distribution with parameters $(a+1, b+1)$. Then

$$F_{H|S_n}(u|k) = \frac{A\int_0^u t^{a+k}(1 - t)^{b+n-k}\,dt}{A\int_0^1 t^{a+k}(1 - t)^{b+n-k}\,dt} \text{ , where } A = C(n,k)\frac{(a + b + 1)!}{a!\,b!}$$

$$= \frac{(a + k)!(n + b - k)!}{(n + a + b + 1)!}\int_0^u t^{a+k}(1 - t)^{b+n-k}\,dt \quad 0 \le u \le 1.$$

Thus, the conditional distribution is Beta, with parameters $(a + k + 1, b + n - k + 1)$. From the formula for expectation, we have

$$E[H|S_n = k] = \frac{a + k + 1}{a + b + n + 2}.$$

It should be noted that since the common factor $C(n,k)$ in the numerator and denominator of the expression for $F_{H|S_n}$ cancel out, the distribution, given $S_n = k$, is the same as that given any specific sequence having k successes and $n - k$ failures. []

The previous development illustrates how conditional independence in a random sample, rather than total independence, may be utilized to modify estimates of probabilities or other parameters which control population probabilities. But <u>decisions</u> are based both on estimates of probabilities and on costs or rewards associated with actions and outcomes. One is apt to proceed much more cautiously (i.e., to require higher probabilities for favorable outcomes) if costs of failure are high, or to be much more venturesome if rewards for success are great. To provide an analytical basis for decision, one must include some measure or criterion of gain or loss, in order that a "best" course of action may be determined. To illustrate, we consider one of the most commonly used criteria: the <u>mean-squared-error criterion</u>.

Suppose $\{X,H\}$ has joint distribution and it is desired to obtain a "best" estimate of the value of H from an experimentally determined value of X. That is, we wish to determine a function or decision rule d such that $d[X(\omega)]$ is the best estimate of $H(\omega)$. According to the mean-squared-error criterion, we seek a function d for which $E\{[H - d(X)]^2\}$ is a minimum. The following argument shows that the best decision function d is given by

$$d(u) = E[H|X = u] = e(u), \quad \forall \ u \text{ in the range of } X.$$

We note that X may be vector valued, in which case u is a vector. Consider

$$0 \leq E\{[H - d(X)]^2\} = E\{[H - e(X) + e(X) - d(X)]^2\}$$

$$= E\{[H - e(X)]^2\} + E\{[e(X) - d(X)]^2\} + 2E\{[H - e(X)][e(X) - d(X)]\}.$$

Suppose we put $h(X) = e(X) - d(X)$. By CE6), $E[Hh(X)] = E[e(X)h(X)]$, so that the last term above is zero. The first term is fixed. The second term is positive, unless $d(X) = e(X)$ a.s., which is equivalent (by Theorem A1-2) to $d(u) = e(u)$ a.s. $[P_X]$. Hence, this choice of d minimizes the mean-squared error.

The argument above solves the regression problem, in which it is desired to determine the random variable $d(X)$ which is "nearest" to H in the mean-squared sense. The central role of conditional expectation is well known. In fact, some authors begin the study of conditional expectation by designating the conditional expectation of X, given Y, as the random variable $e(Y)$ for which the mean-squared error $E\{[X - e(Y)]^2\}$ is a minimum. Starting from this point, it is possible to show that $e(Y)$ has all the properties of the concept as we have introduced it.

Example D2-b

Returning to the situation presented in Example D2-a, we suppose n items are selected at random from the production lot and tested. Of these, k

meet specifications. What is the best estimate, in the mean-squared sense, of the probability that any item selected will meet specifications?

SOLUTION AND DISCUSSION.

By the development above, if the prior distribution for H is Beta $(a+1, b+1)$, the best estimator for H, given S_n, is

$$E[H|S_n] = \frac{a + S_n + 1}{a + b + n + 2} \quad \text{as compared with} \quad E[H] = \frac{a + 1}{a + b + 2}.$$

The rule is: count the number of successes in the n units tested, add $a + 1$, and divide by $a + b + n + 2$.

Suppose no prior information about f_H is available; we should use $a = b = 0$.

Suppose, further, that in a test of 10 items, 8 meet specifications. Then

$$E[H|S_{10} = 8] = \frac{8 + 1}{10 + 2} = 9/12 = 3/4 \quad \text{as compared with} \quad E[H] = 1/2.$$

If the prior distribution were Beta with $a = 2$, $b = 1$, then

$$E[H|S_{10} = 8] = \frac{8 + 3}{10 + 2} = 11/15 \approx 0.7333 \quad \text{as compared with} \quad E[H] = 3/5.$$

The conditional distribution for H, given $S_n = k$, is Beta $(a+k+1, b+n-k+1)$. The conditional variance is

$$\text{Var}[H|S_n = k] = \frac{(a + k + 1)(b + n - k + 1)}{(a + b + n + 2)^2(a + b + n + 3)}.$$

For $a = b = 0$, $n = 10$, $k = 8$,

$$\text{Var}[H|S_{10} = 8] = (9 \times 3)/(12^2 \times 13) = 3/208 \approx 0.0144.$$

For $a = 2$, $b = 1$, $n = 10$, $k = 8$,

$$\text{Var}[H|S_{10} = 8] = (11 \times 4)/(15^2 \times 16) = 11/900 \approx 0.0122.$$

The prior information, with its approximate location and indication of variance, gives rise to a somewhat smaller variance on the conditional distribution. []

For a more general discussion of the problem of Bayesian estimation, as this procedure is called, see Mood, Graybill, and Boes [1974], Chap VII, Sec 7. Although they do not employ the term conditional independence, they assume it by virtue of assuming the product rule for conditional

densities. They consider other measures of distance or "loss", and relate the results to the results of other estimation procedures commonly employed in modern statistics.

3. A one-stage Bayesian decision model

The transition from inference (i.e., determining the most likely alternative) to decision (determining the course of action to be selected) leads to the notion of gain or loss. In order to move beyond a purely mathematical criterion such as mean-squared error, we introduce the notion of a loss function. The loss function is usually expressed in terms of some symbol of value, such as monetary units. But its specification may require quite subtle and subjective judgments of "utility" or worth. In order to be objective, the decision analyst must obtain from the decision maker enough information to determine a loss function whose value depends upon the course of action chosen and the resultant outcome of this action. To set up a model of a typical decision process, we suppose:

i) There is a set of possible actions available to the decision maker. Action a is a member of the set A of possible actions.

ii) There is a set of possible outcomes which may result from the action. Because there is uncertainty about which consequence will materialize, we represent the outcome as the value of an outcome random variable (or random vector): $y = Y(\omega)$.

iii) The distribution of the outcome random variable Y is determined by a state of nature. This is often expressed as a parameter (possibly vector-valued). Since there is uncertainty about the state of nature, the parameter itself is modeled as the value of a parameter random variable: $u = H(\omega)$.

iv) It may be possible to experiment in order to obtain some information about the state of nature. The result of the experiment is the value of a test random variable: $x = X(\omega)$. Both Y and X are jointly distributed with the parameter random variable H.

v) A loss function L is determined. L(a,y) is the loss when action a is taken and outcome y is experienced (a gain is a negative loss). The usual objective is to minimize the expected loss.

vi) If experimentation is utilized, a decision rule (or strategy) is determined, to indicate the action to be taken for each possible observed value of the test random variable. In practice, the value of the decision rule may be determined for only the specific experimental result observed.

We consider two cases.

a) Without experimentation.

Assume $\{Y,H\}$ have joint distribution. Let $\ell(a,u) = E[L(a,Y)|H = u]$. This is sometimes known as the risk function. The objective is to select action a to minimize $R(a) = E[L(a,Y)] = E\{E[L(a,Y)|H]\}$ $= E[\ell(a,H)]$. In some problems, $Y = H$, so that $\ell(a,u) = L(a,u)$. In the case of no experimentation, no conditional independence assumptions are needed.

Example D3-a

A merchant plans to stock an item. The demand over a six-week period is assumed to be a random quantity having the Poisson distribution, with parameter λ. The parameter value is not known, but on the basis of past experience the merchant assumes λ to be the value of a random variable H with possible values $\{15, 20, 25\}$ taken on with probabilities $\{1/4, 1/2, 1/4\}$, respectively. The merchandise may be ordered in lots of 10. The merchant contemplates ordering either 10, 20, or 30 units. He can buy at a cost of $c = \$7$ per unit; he can sell at a price $u = \$10$ per unit. At the end of six weeks, he can return the unsold items for a

net recovery of $r = \$3$ per unit, so that he loses $c - r = \$4$ per unsold unit. He considers that he has lost $(u - c)/2 = \$1.50$ per missed sale. From a Bayesian point of view, how many units should he order?

SOLUTION

The set of possible actions is $A = \{10, 20, 30\}$. Let Y be the random variable whose value is the demand in the six-week period (the outcome random variable). The conditional distribution of Y, given $H = \lambda$, is assumed to be Poisson (λ). The loss function L is given by

$$L(a,y) = \begin{cases} -(u-c)y + (c-r)(a-y) = 4a - 7y & \text{for } y \leq a \\ -(u-c)a + \frac{u-c}{2}(y-a) = -4.5a + 1.5y & \text{for } y > a. \end{cases}$$

If we set $B = \{Y \leq a\}$, we may then write

$$L(a,Y) = I_B(4a - 7Y) + (1 - I_B)(1.5Y - 4.5a) = 1.5Y - 4.5a - 8.5I_B(Y - a).$$

Now $\ell(a,\lambda) = E[L(a,Y)|H = \lambda]$

$$= 1.5E[Y|H = \lambda] - 4.5a + 8.5aP(Y \leq a|H = \lambda) - 8.5E[I_B Y|H = \lambda].$$

We may express

$$E[I_B Y|H = \lambda] = \sum_{k=0}^{a} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=0}^{a-1} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda P(Y \leq a-1|H = \lambda).$$

Hence

$$\ell(a,\lambda) = \lambda[1.5 - 8.5P(Y \leq a-1|H = \lambda)] - a[4.5 - 8.5P(Y \leq a|H = \lambda)].$$

Using a table of cumulative or summed Poisson distribution for appropriate values of λ, we may establish the following table of values for $\ell(a,\lambda)$.

$a =$	10	20	30	
$\lambda = 15$	- 21.3	- 23.2	+ 15.0	
20	- 14.9	- 44.9	- 19.7	$\ell(a,\lambda)$
25	- 7.4	- 49.5	- 51.3	

Now $R(a) = E[L(a,Y)] = E[\ell(a,H)]$ has values:

$$R(10) = \frac{1}{4}[-21.3 - 14.9 \times 2 - 7.4] = -14.6; \quad R(20) = -40.6; \quad \text{and}$$

$R(30) = -18.9$.

The optimum action, corresponding to the minimum expected loss, is $a = 20$. []

b) With experimentation

Assume $\{X,Y,H\}$ has a joint distribution. A decision is made on the basis of the experimental data (i.e., on the basis of the observed value of X). The problem is to determine the optimum decision function d^* which designates the optimum action $d^*(x)$ when the test random variable X has value x. Thus, d^* is the decision function which minimizes the Bayesian risk $B(d) = E[L(d(X),Y)]$.

The problem may be formulated in a useful way as follows. By CE1b), $B(d) = E\{E[L(d(X),Y)|X]\}$. If we set $R(a,x) = E[L(a,Y)|X = x]$, then by CE10), $R(d(x),x) = E[L(d(x),Y)|X = x] = E[L(d(X),Y)|X = x]$. Thus, $R(d(X),X) = E[L(d(X),Y)|X]$, so that $B(d) = E[R(d(X),X)]$. For each x in the range of X, let $d^*(x)$ be the action for which $R(d^*(x),x)$ is a minimum. Then $B(d^*) = E[R(d^*(X),X)] \leq E[R(d(X),X)] = B(d)$, for all possible decision functions d.

In the usual situation, the result of experimentation does not affect operationally the outcome following the action. The experimental evidence may be in the form of previously available data. The result of a given action is not influenced by whether or not the decision maker obtains the experimental data. What does affect the outcome following an action is the value of the"state of nature" parameter. Thus, it is appropriate to assume the pair $\{X,Y\}$ is conditionally independent, given H. We utilize this as follows.

1) If X is discrete, we may use CI10b) to assert

$$R(a,x) = E[L(a,Y)|X = x] = E\{E[I_{\{x\}}(X)|H]E[L(a,Y)|H]\}/E\{E[I_{\{x\}}(X)|H]\},$$

where $E[I_{\{x\}}(X)|H = u] = P(X = x|H = u) = p_{X|H}(x|u)$ and

$$E[L(a,Y)|H] = \ell(a,H).$$

Hence,

$$R(a,x) = \int \ell(a,u)\, p_{X|H}(x|u)\, dF_H(u) \,/P(X = x).$$

2) If X is absolutely continuous,

$$R(a,x) = E[L(a,Y)|X = x] = E\{E[L(a,Y)|H]|X = x\} \qquad \text{by CI8)}$$

$$= E[\ell(a,H)|X = x] = \int \ell(a,u)\, dF_{H|X}(u|x).$$

We may use Bayes' theorem for the conditional distribution (see Sec. C6) to determine $F_{H|X}$.

Example D3-b

Suppose in Example D3-a the merchant recalls that he made a similar order for a corresponding period the previous year. If X is the random variable whose value represents the demand for that period, an observation of the value for that period should provide some indication of the state of the market for the period. If there is reason to believe that the state of the market has not changed appreciably, this information should be useful for the present decision. Enough time has elapsed that sales in the previous period should not influence directly sales in the current period. Therefore, it seems reasonable to assume that $\{X,Y\}$ is conditionally independent, given H (the value of which indicates the general state of the market). A check of the previous sales records shows that demand was for 24 units. Under these assumptions and with these data, the task is to select $a = d^*(24)$ to minimize $R(a,24) = \sum_{\lambda} \ell(a,\lambda)p_{X|H}(24|\lambda)p_H(\lambda)/P(X = 24)$. Values of $\ell(a,\lambda)$ are tabulated in the solution of Example D3-a. Under the assumed conditions, we may reasonably suppose $p_{X|H} = p_{Y|H}$. From tables of the Poisson distribution, we obtain values of $p_{X|H}(24|\lambda)$, from which we determine

$$p_X(24) = p_{X|H}(24|15)p_H(15) + p_{X|H}(24|20)p_H(20) + p_{X|H}(24|25)p_H(25)$$

$$= \frac{1}{4}\,[0.0083 + 0.0557 \times 2 + 0.0795] = 0.050$$

and

$$R(10,24) = \frac{-21.3 \times 0.0083 - 14.9 \times 2 \times 0.0557 - 7.4 \times 0.0795}{4 \times 0.050} = -12.13\,.$$

The values $R(20,24) = -45.8$ and $R(30,24) = -30.9$ may be calculated in similar fashion. Once more, the indicated optimum action is to order 20 units. In spite of the fact that the previous demand went beyond 20 units, the best bet is to order 20 units and risk the loss of some sales. []

4. A dynamic programming example

The following example of a multistage decision process is presented in Gaver and Thompson [1973], p. 392 ff. Our discussion displays the role of a conditional independence assumption, which seems to be both appropriate and necessary.

Example D4-a

A company is offered two investment opportunities, which we designate "risk" and "safe".

1) Risk. Either make gain g in a given period or earn nothing. Probability of success is unknown, but constant, over the total time considered.

2) Safe. Certain to make gain s in the given period.

Gains in successive periods are independent, given a fixed probability of success. A choice is made at the beginning of each time period, with negligible cost for switching from one investment to the other. The objective is to maximize expected gain over N time periods.

SOLUTION.

The probability of success is unknown; we suppose that it is the value of a state-of-nature random variable H. A prior density f_H (or distribution function F_H) is assumed. To obtain further information, the company must experiment by making the risky investment. Suppose I_k is the indicator function for success in the kth risk period (i.e., $I_k(\omega) = 1$ iff the risk pays off on the kth trial). The gain during that period is gI_k. We assume the class $\{I_k: 1 \leq k \leq N\}$ is identically distributed, conditionally independent, given H, with $E[I_k|H = t] = P[I_k = 1|H = t) = t$. Suppose n risks have been taken; let S_n be the random variable which counts the

number of successes-- i.e., $S_n = I_1 + I_2 + \ldots + I_n$. The succession of choices to take the risky alternative constitutes a Bernoulli sequence, with conditional independence, given the parameter random variable H. In Sec D-2, we establish an expression for

$$p(n,k) = E[I_{n+1} | S_n = k] = E[H | S_n = k], \text{ when } H \text{ has the Beta distribution}$$

If there is no basis for assigning a given prior distribution for H, we assign the uniform distribution. According to the results in Example D2-b, we have

$$p(n,k) = \frac{k+1}{n+2}.$$

To develop a strategy based on optimum expected gain, we utilize the backward induction procedure of dynamic programming. Consider the beginning of the jth period. If n risks have been taken before stage j, then there is an "optimum-path" gain random variable $G_{n,j} = f_{n,j}(S_n, I_{n+1}, \ldots, I_N)$. At most N risks will be taken, but not necessarily this many. It is convenient to use a decision tree to keep account of the alternatives (see Fig. D4-1).

Suppose $S_n = k$. The decision rule is risk iff

$$E[gI_{n+1} + G_{n+1,j+1} | S_n = k] \geq s + E[G_{n,j+1} | S_n = k].$$

We wish to obtain an expression for $G_{n,j}$. Consider the set

$$M = \{k: E[gI_{n+1} + G_{n+1,j+1} | S_n = k] \geq s + E[G_{n,j+1} | S_n = k]\}.$$

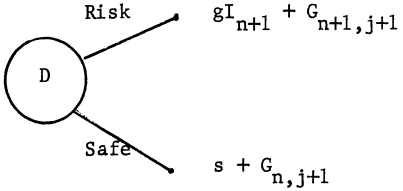


Figure D4-1. Alternatives at decision nodes.

Then

$$G_{n,j} = I_M(S_n)(gI_{n+1} + G_{n+1,j+1}) + [1 - I_M(S_n)](s + G_{n,j+1}) \quad \text{and}$$

$$E[G_{n,j}|S_n = k] = \max \{E[gI_{n+1} + G_{n+1,j+1}|S_n = k],\ s + E[G_{n,j+1}|S_n = k]\}.$$

Since $E[X|A] = E[X|AB]P(B|A) + E[X|AB^c]P(B^c|A)$, $E[gI_{n+1}|S_n = k,\ I_{n+1} = 1] = g$,

and $P(I_{n+1} = 1|S_n = k) = p(n,k)$, we obtain

$$E[gI_{n+1} + G_{n+1,j+1}|S_n = k]$$

$$= \{g + E[G_{n+1,j+1}|S_{n+1} = k+1]\}p(n,k) + E[G_{n+1,j+1}|S_n = k][1 - p(n,k)]$$

$$= [g + \varphi_{j+1}(n+1,k+1)]p(n,k) + \varphi_{j+1}(n+1,k)[1 - p(n,k)],$$

where $\varphi_j(n,k) = E[G_{n,j}|S_n = k]$. We may formulate the decision rule as follows:

$$\varphi_j(n,k) = \max \{[g + \varphi_{j+1}(n+1,k+1)]p(n,k) + \varphi_{j+1}(n+1,k)[1 - p(n,k)],\ s + \varphi_{j+1}(n,k)$$

with $\varphi_{N+1}(n,k) = 0$.

To see how the procedure goes, let $g = 5/2$, $s = 1$, $N = 2$, $f_H(t) = 1$ on $[0,1]$, so that $p(n,k) = \frac{k+1}{n+2}$. Refer to Figure D4-2 for situations at decision nodes.

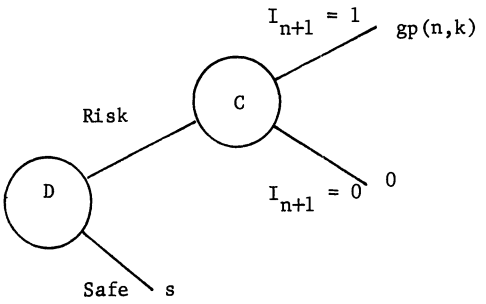
<u>At the final decision node</u>, $j = N = 2$, and $(n,k) = (0,0)$, $(1,0)$, or $(1,1)$

Determine $\varphi_2(0,0)$, $\varphi_2(1,0)$, $\varphi_2(1,1)$ and the optimum action in each case.

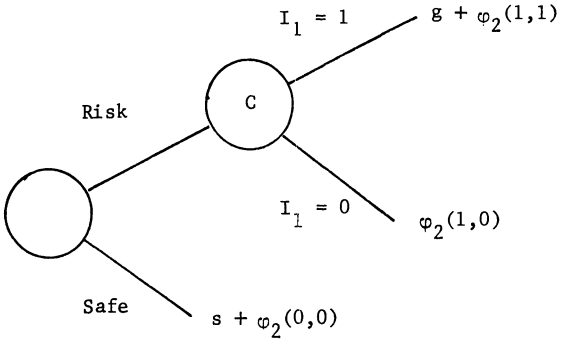
$p(0,0) = 1/2$, $p(1,0) = 1/3$, $p(1,1) = 2/3$

$$\varphi_2(0,0) = \max \{\tfrac{1}{2}g + 0,\ s\} = \max \{\tfrac{5}{4},\ 1\} = 5/4 \qquad \text{(risk)}$$

$$\varphi_2(1,0) = \max \{\tfrac{1}{3}g + 0, s\} = \max \{\tfrac{5}{6},\ 1\} = 1 \qquad \text{(safe)}$$



a) At the final decision node



b) At the initial decision node

Figure D4-2. Decision nodes for Example D4-a .

$$\varphi_2(1,1) = \max\{\tfrac{2}{3}g + 0,\ s\} = \max\{\tfrac{5}{3},\ 1\} = 5/3 \qquad \text{(risk)}.$$

At the initial decision node, $j = 1$ and $(n,k) = (0,0)$

$$\varphi_1(0,0) = \max\{[g + \varphi_2(1,1)]\tfrac{1}{2} + \varphi_2(1,0)\tfrac{1}{2},\ s + \varphi_2(0,0)\}$$

$$= \max\{[5/2 + 5/3]\tfrac{1}{2} + 1/2,\ 1 + 5/4\}$$

$$= \max\{31/12,\ 27/12\} = 31/12 \qquad \text{(risk)}.$$

The indicated strategy is:

First decision: Risk $\sim \varphi_1(0,0) = 31/12$

Second decision: If first risk is successful $\sim \varphi_2(1,1) \Rightarrow$ Risk.

If first risk unsuccessful $\sim \varphi_2(1,0) \Rightarrow$ Safe.

The expected gain from this strategy is $\varphi_1(0,0) = 31/12 \approx 2.58$. []

5. Proofs of the basic properties

To establish the equivalence of Properties CI1) through CI4), we show CI1) $\Leftrightarrow$ CI2) $\Rightarrow$ CI3) $\Rightarrow$ CI4) $\Rightarrow$ CI2). To simplify writing, we drop the "a.s." in the step-by-step arguments.

CI1) $\Rightarrow$ CI2)

$$E\{I_N(Y)E[I_M(X)|Z]\,|Z\} = E[I_M(X)|Z]E[I_N(Y)|Z] \quad \text{by CE8)}$$
$$= E[I_M(X)I_N(Y)|Z] \quad \text{by CI1)}$$
$$= E\{E[I_M(X)I_N(Y)|Z,Y]\,|Z\} \quad \text{by CE9)}$$
$$= E\{I_N(Y)E[I_M(X)|Z,Y]\,|Z\} \quad \text{by CE8).}$$

Now

$$E(I_Q(Z)E\{I_N(Y)E[I_M(X)|Z]\,|Z\})$$
$$= E(I_Q(Z)I_N(Y)E[I_M(X)|Z]) \quad \forall \text{ Borel } Q \quad \text{by CE1).}$$

A similar expression holds for all Borel Q with $E[I_M(X)|Z]$ replaced by $E[I_M(X)|Z,Y]$. We thus have

$E\{I_Q(Z)I_N(Y)E[I_M(X)|Z]\} = E\{I_Q(Z)I_N(Y)E[I_M(X)|Z,Y]\}$ for all Borel sets N, Q on the codomains of Y, Z, respectively. By E6b), we may assert

$$E[I_M(X)|Z] = e_1(Z,Y) = e_2(Z,Y) = E[I_M(X)|Z,Y] \quad \text{a.s.}$$

CI2) $\Rightarrow$ CI1)

$$E[I_M(X)I_N(Y)|Z] = E\{E[I_M(X)I_N(Y)|Z,Y]\,|Z\} \quad \text{by CE9)}$$
$$= E\{I_N(Y)E[I_M(X)|Z,Y]\,|Z\} \quad \text{by CE8)}$$
$$= E\{I_N(Y)E[I_M(X)|Z]\,|Z\} \quad \text{by CI2)}$$
$$= E[I_M(X)|Z]E[I_N(Y)|Z] \quad \text{by CE8).}$$

CI2) ⇒ CI3)

$$E[I_M(X)I_Q(Z)|Z,Y] = I_Q(Z)E[I_M(X)|Z,Y] \quad \text{by CE8)}$$
$$= I_Q(Z)E[I_M(X)|Z] \quad \text{by CI2)}$$
$$= E[I_M(X)I_Q(Z)|Z] \quad \text{by CE8).}$$

CI3) ⇒ CI4)

$$E[I_M(X)I_Q(Z)|Y] = E\{E[I_M(X)I_Q(Z)|Z,Y]|Y\} \quad \text{by CE9)}$$
$$= E\{E[I_M(X)I_Q(Z)|Z]|Y\} \quad \text{by CI3).}$$

CI4) ⇒ CI2)

$$E\{E[I_M(X)I_Q(Z)|Z]|Y\} = E[I_M(X)I_Q(Z)|Y] \quad \text{by CI4)}$$
$$= E\{E[I_M(X)I_Q(Z)|Z,Y]|Y\} \quad \text{by CE9).}$$

This ensures that for all Borel N on the codomain of Y

$$E\{I_N(Y)E[I_M(X)I_Q(Z)|Z]\} = E\{I_N(Y)E[I_M(X)I_Q(Z)|Z,Y]\} \quad \text{by CE1).}$$

But this, in turn, ensures that

$$E\{I_N(Y)I_Q(Z)E[I_M(X)|Z]\} = E\{I_N(Y)I_Q(Z)E[I_M(X)|Z,Y]\} \quad \text{by CE8).}$$

By E6b), we must have

$E[I_M(X)|Z] = E[I_M(X)|Z,Y]$ a.s., which is CI2). []

We wish to establish next the equivalence of CI5) through CI7) to the propositions above. It is apparent by the special-case relationship that CI5) ⇒ CI1), CI7) ⇒ CI6) ⇒ CI2), and CI8) ⇒ CI4). Extension of CI1) to CI5) may be done by a "standard argument" based on linearity, monotonicity, monotone convergence, and approximation by step functions. Extension of CI3) to CI7) may be achieved by an argument similar to that sketched in the discussion of the proof of CE10), plus a "standard argument." A similar approach serves to extend CI4) to CI8).

Before proving CI9), we obtain a lemma useful here and elsewhere.

Lemma D5-1

If $E[g(W)|V,U] = E[g(W)|V]$ a.s. and $Z = h(U)$, with h Borel,

then $E[g(W)|V,Z] = E[g(W)|V]$ a.s.

PROOF

The random vector $(V,Z) = (V,h(U))$ is a Borel function of (V,U). Hence,

$E[g(W)|V,Z] = E\{E[g(W)|V,U]|V,Z\}$ a.s. by CE9)

$= E\{E[g(W)|V]|V,Z\}$ a.s. by hypothesis

$= E[g(W)|V]$ a.s. by CE9a). []

PROOF OF CI9)

For any Borel function g,

$E[g(X)|Z] = E[g(X)|Z,Y]$ a.s. by CI6)

$= E[g(X)|Z,V]$ a.s. by Lemma D5-1.

Hence, $\{X,V\}$ is conditionally independent, given Z by CI6).

For any Borel function r,

$E[r(V)|Z] = E[r(V)|Z,X]$ a.s. by CI6)

$= E[r(V)|Z,U]$ a.s. by Lemma D5-1.

Hence, $\{U,V\}$ is conditionally independent, given Z by CI6). []

PROOF OF CI10)

a) $E[g(X)h(Y)] = E\{E[g(X)h(Y)|Z]\}$ by CE1b)

$= E\{E[g(X)|Z]E[h(Y)|Z]\}$ by CI5)

$= E[e_1(Z)e_2(Z)]$ (notational change).

b) $E[g(X)|Y \in N]P(Y \in N) = E[I_N(Y)g(X)]$ by CE1a)

$= E\{E[I_N(Y)|Z]E[g(X)|Z]\}$ by part a). []

PROOF OF CI11)

Given that $\{Y, (X,Z)\}$ is independent.

$P(X \in M, Y \in N, Z \in Q) = E[I_M(X)I_N(Y)I_Q(Z)]$ by E1a)

$= E\{E[I_M(X)I_N(Y)I_Q(Z)|Z]\}$ by CE1b)

$= E\{I_Q(Z)E[I_M(X)I_N(Y)|Z]\}$ by CE8).

Also,

$P(X \in M, Y \in N, Z \in Q) = P(Y \in N)P(X \in M, Z \in Q)$ by independence

$= E[I_N(Y)]E[I_M(X)I_Q(Z)]$ by E1a)

$= E[I_N(Y)]E\{I_Q(Z)E[I_M(X)|Z]\}$ by CE1)

$= E\{I_Q(Z)E[I_N(Y)]E[I_M(X)|Z]\}$ by E2)

$= E\{I_Q(Z)E[I_N(Y)|Z]E[I_M(X)|Z]\}$ by CE5).

Equating the last expressions in each series of inequalities, by E6) we conclude that $E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z]E[I_N(Y)|Z]$ a.s. []

PROOF OF CI12)

As in the proof of CE10), it is sufficient to show the proposition holds for $g = I_{M\times N} = I_M I_N$.

$E[I_M(X)I_N(Y)|Y = u, Z = v] = I_N(u)E[I_M(X)|Y = u, Z = v]$ by CE8)

$= I_N(u)E[I_M(X)|Z = v]$ by CI6)

$= E[I_M(X)I_N(u)|Z = v]$ a.s. $[P_{YZ}]$ by CE2). []

6. Problems

D-1 Show that if $\{X,(Y,Z)\}$ is independent, then

$$E[g(X)h(Y)|Z] = E[g(X)]E[h(Y)|Z] \quad \text{a.s.}$$

D-2 Let $\{X_i : 1 \le i \le n\}$ be a random sample, given H. Determine the best mean-square estimate for H, given $W = (X_1, X_2, \ldots, X_n)$ for each of the following cases:

i) X is delayed exponential: $f_{X|H}(t|u) = e^{-(t-u)}$ for $t \ge u$, and H is exponential (1): $f_H(u) = e^{-u}$ for $u \ge 0$

ii) X is Poisson (u): $p_{X|H}(k|u) = e^{-u}\frac{u^k}{k!}$ $k = 0, 1, 2, \ldots$, and H is gamma (m,λ): $f_H(u) = \lambda^m u^{m-1} e^{-\lambda u}/(m-1)!$ $u \ge 0$, $m > 0$, $\lambda > 0$

iii) X is geometric (u): $p_{X|H}(k|u) = u(1-u)^k$ $k = 0, 1, 2, \ldots$, and H is uniform $[0,1]$.

D-3 In Example D2-b, suppose $a = 7$, $b = 3$. Compare the prior density for H and the quantities $E[H|S_{10} = 8]$ and $\mathrm{Var}[H|S_{10} = 8]$ with those for the case $a = 2$, $b = 1$, as in the example.

D-4 Consider the demand random variable of problem C-13:
$D = \sum_{i=1}^{N} X_i = \sum_{n=0}^{\infty} I_{\{n\}}(N)Y_n$, where $Y_0 = 0$, $Y_n = X_1 + X_2 + \ldots + X_n$, $n \ge 1$
Suppose $\{N,(H,X_1,X_2, \ldots, X_n)\}$ is independent for each $n \ge 1$, and $E[X_i|H = u] = e(u)$, invariant with i. Show that $E[D|H] = E[N]e(H)$.

D-5 It is desired to study the waiting time for the arrival of an ambulance after reporting an accident (see Scott, et al, [1978]). Direct statistical data are difficult to obtain. Suppose we consider the random variables

N = number of ambulances in service (integer-valued)

D = distance traveled by dispatched ambulance

V = average velocity of the ambulance for the trip.

By considering the geometry of the deployment scheme, it is possible to make reasonable assumptions about $P(D \le t|N = n)$. Also, it is

possible to make reasonable assumptions, on the basis of statistical data, for the distribution of V and the distribution of N. We have $W = D/V$, where W is the random variable whose value is the waiting time. Then $P(W \leq t) = E[I_Q(D,V)]$, where $Q = \{(u,v): u \leq vt\}$.

a) Show that if $\{V,(D,N)\}$ is independent, then

$P(W \leq t|N = n) = \int P(D \leq vt|N = n)\, dF_V(v)$

Suggestion. Use CE1b), CI11), CI12).

b) Under the conditions for part a) and the assumptions

i) $P(D \leq s|N = n) = as$, $0 \leq s \leq 1/a$, where $a^2 = n\pi/A$

ii) V is uniform $[15,25]$

where A is the area served, in square miles, D is distance in miles, and V is velocity, in miles per hour.

c) Repeat part b) with i) replaced by

i') $P(D \leq s|N = n) = 1 - e^{-as}$, $s \geq 0$, $a^2 = n\pi/A$.

D-6 In Example D3-b, suppose the previous demand was 26 units. What is the optimum action?

D-7 An electronic game is played as follows. A probability of success in a sequence of Bernoulli trials is selected at random. A player is allowed to observe the result of m trials. He is then to guess the the number of successes in the next n trials. If he guesses within one of the actual number of successes, he gains one dollar (loses -1); if his guess misses by two or more, he loses one dollar. Suppose $m = 3$, $n = 10$; on the trial run there are two out of three successes. What number should he then guess to minimize his expected loss? Let

X = number of successes in m on the trial run

Y = number of successes in n on the pay run

H = parameter random variable.

Then X is binomial (m,u), given $H = u$

Y is binomial (n,u), given $H = u$

H is uniform on $[0,1]$

and $L(a,y) = \begin{cases} -1 & \text{for } |a - y| \leq 1 \\ 1 & \text{for } |a - y| > 1 \end{cases}$

D-8 In Example D4-a, determine the optimum strategy for $g = 5/2$, $s = 1$, $N = 3$, H uniform on $[0,1]$.

E. Markov Processes and Conditional Independence

E. MARKOV PROCESSES AND CONDITIONAL INDEPENDENCE

E. Markov processes and conditional independence

1. Discrete-parameter Markov processes

The notion of conditional independence has been utilized extensively in advanced treatments of Markov processes. Such processes appear as models of processes without "memory." The "future" is conditioned only by the"present" and not by the manner in which the present state is reached. The past thus affects the future only as it influences the present. We wish to make the connection between the usual introductory treatment and the more advanced point of view, which is not only mathematically powerful but intuitively helpful in displaying the essential character of Markov processes. For a recent introductory treatment utilizing conditional independence, see Çinlar [1975].

Many elementary textbooks include a treatment of Markov processes with discrete parameter and finite, or at most countably infinite, state space. Suppose we have a sequence $\{X_n: 0 \leq n\}$ of random variables, each with range $\{0, 1, 2, \ldots, N\}$. Thus, the parameter set is $T = \{0, 1, 2, \ldots\}$ and the state space is $S = \{0, 1, 2, \ldots, N\}$. The Markov property is expressed by the condition

M) $P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0)$

$= P(X_{t+1} = j | X_t = i) = p_{ij}(t)$

for all $t \geq 1$, all $(i,j) \in S^2$, all $(i_0, i_1, \ldots, i_{t-1}) \in S^t$.

The quantities $p_{ij}(t)$ are called the transition probabilities. In the important case of stationary (or homogeneous) transition probabilities, we have $p_{ij}(t) = p_{ij}$, invariant with t. In this case, analysis is largely algebraic, with the transition matrix $\mathbf{P} = [p_{ij}]$ playing a central role.

The fundamental notion of the Markov property M) is that the past does not condition the future, except as it influences the present. We can

give the Markov property M) an alternative formulation which emphasizes the conditional independence of past and future, given the present, without restriction to discrete state space or to stationary transition probabilities. To aid in formulating this condition, we introduce the following notation. If S is the state space, then

S^k = set of all k-tuples of elements of state space S

$U_s = (X_0, X_1, \ldots, X_s)$ $\qquad U_s: \Omega \to S^{s+1}$

$V_{s,t} = (X_s, X_{s+1}, \ldots, X_t)$ $\qquad V_{s,t}: \Omega \to S^{t-s+1}$ $\qquad s \leq t$

$W_{t,u} = (X_t, X_{t+1}, \ldots, X_u)$ $\qquad W_{t,u}: \Omega \to S^{u-t+1}$ $\qquad t \leq u.$

We indicate by U_s^* a random vector whose coordinates consist of a subset (in natural order) of the coordinates of U_s, and similarly for $V_{s,t}^*$ and $W_{t,u}^*$. Then U_s^*, $V_{s,t}^*$, and $W_{t,u}^*$ are continuous, hence Borel, functions of U_s, $V_{s,t}$, and $W_{t,u}$, respectively. When we write a function $g(U_s)$, $h(V_{s,t})$, etc., we suppose g, h, etc. are real-valued Borel functions such that $E[g(U_s)]$, $E[h(V_{s,t})]$, etc. are all finite.

If t represents the "present", then U_{t-1} represents the"past behavior" of the process and $W_{t+1,u}$ represents the behavior of the process for a "finite future." We sometimes consider an "extended present", represented by $V_{s,t}$, $s < t$.

In this notation, the Markov property M) is equivalent to

$$P(X_{t+1} \in M | X_t = u, U_{t-1} = v) = P(X_{t+1} \in M | X_t = u)$$

$\forall\, t \geq 1$, $\forall$Borel sets $M \subset S$, $\forall\, u \in S$, $\forall\, v \in S^t$

which is equivalent to

M) $E[I_M(X_{t+1}) | X_t, U_{t-1}] = E[I_M(X_{t+1}) | X_t]$ a.s. $\forall\, t \geq 1$, $\forall$ Borel $M \subset S$.

Reference to CI2) shows property M) is equivalent to

M') $\{X_{t+1}, U_{t-1}\}$ is conditionally independent, given X_t, $\forall\, t \geq 1$.

DEFINITION. The process $\{X_t: t \in T\}$ is Markov iff M') holds.

Use of CI1) through CI8) provides a number of alternative formulations of the basic condition M). It is sometimes desirable to remove the restriction to the immediate future. This can be done (and more), as the following theorem shows.

Theorem E1-1

A process $\{X_t: t \in T\}$, $T = \{0, 1, 2, \ldots\}$ is Markov iff

M'') $\{W^*_{t+1,t+n}, U^*_{s-1}\}$ is conditionally independent, given any finite extended present $V_{s,t}$, $1 \leq s \leq t$, any $n \geq 1$, any $W^*_{t+1,t+n}$, U^*_{s-1}.

A proof is given in Sec E5. []

To see how the idea of conditional independence is an aid to modeling, we consider several examples.

Example E1-a One-dimensional random walk

A number of physical and behavioral situations can be represented schematically as "random walks." A particle is positioned on a line. At discrete instants of time $t_1, t_2, \ldots$, the particle moves an amount represented by the values of the random variables $Y_1, Y_2, \ldots$, respectively. Positive values indicate movements in one direction and negative values indicate movements in the opposite direction. The position after the nth move is $X_n = Y_1 + Y_2 + \ldots + Y_n$ (we take $X_0 = 0$). If we can assume the class $\{Y_i: 1 \leq i\}$ is independent, then $X_{n+1} = X_n + Y_{n+1}$, with $\{Y_{n+1}, (U_{n-1}, X_n)\}$ independent for all $n \geq 0$. Since the position at time t_{n+1} is affected by the past behavior only as that behavior affects the present position X_n (at time t_n), it seems reasonable to suppose that the Markov condition holds. []

Example E1-b A class of branching processes

Consider a population consisting of "individuals" able to produce new individuals of the same kind. We suppose the production of new individuals occurs at specific instants for a whole "generation." To avoid the complication of a possibly infinite population in some generation, we suppose a mechanism operates to limit the total population to M individuals at any time. Let X_0 be the original population and suppose the number of individuals produced by each individual in a given generation is a random variable. Let Z_{in} be the random variable whose value is the number of individuals produced by the ith member of the nth generation. If $Z_{in} = 0$, that individual does not survive; if $Z_{in} = 1$, either that individual survives and produces no offspring or does not survive and produces one offspring. If X_n is the number of individuals in the nth generation, then

$$X_{n+1} = \min \{ M, \sum_{i=1}^{X_n} Z_{in} \} = g(X_n, Y_{n+1}), \text{ where } Y_{n+1} = (Z_{1n}, Z_{2n}, \cdots, Z_{Mn}).$$

If $\{Z_{in}: 1 \leq i \leq M, 0 \leq n < \infty\}$ is an independent class, then $\{Y_{n+1}, (U_{n-1}, X_n)\}$ is an independent pair for any $n \geq 0$. Again we have a situation in which past behavior affects the future only as it affects the present. It seems reasonable to suppose the process $\{X_n: 0 \leq n\}$ is Markov. []

Example E1-c An inventory problem

A store uses an (m,M) inventory policy for a certain item. This means:

If the stock at the end of a period is less than m, "order up" to M

If the stock at the end of the period is as much as m, do not order.

Suppose the merchant begins the first period with a stock of M units. Let X_n be the stock at the end of the nth period $(X_0 = M)$. If the demand during the nth period is D_n, then

$$X_{n+1} = \begin{cases} \max\{(M - D_{n+1}), 0\} & \text{if } 0 \le X_n < m \\ \max\{(X_n - D_{n+1}), 0\} & \text{if } m \le X_n \le M \end{cases} = g(X_n, D_{n+1}).$$

If we suppose $\{D_n: 1 \le n\}$ is an independent class, then we have $\{D_{n+1}, (U_{n-1}, X_n)\}$ is an independent pair for each $n \ge 0$. Once more it seems the past and future should be conditionally independent, given the present. []

Each of these examples provides a special case of the following

Theorem E1-2

Suppose $\{Y_n: 1 \le n\}$ is an independent class of random vectors. Set $X_0 = c$ (a constant) and for $n \ge 0$ let $X_{n+1} = g_{n+1}(X_n, Y_{n+1})$. Then the process $\{X_n: 0 \le n\}$ is Markov and

$$P(X_{n+1} \in Q | X_n = u) = P[g_{n+1}(u, Y_{n+1}) \in Q] \quad \forall\, n \ge 0, \ \forall\, u \in S, \ \forall \text{ Borel set } Q$$

PROOF

$U_k = (X_0, X_1, \ldots, X_k) = h_k(Y_1, Y_2, \ldots, Y_k)$, $1 \le k \le n$. Thus, $\{Y_{n+1}, (U_{n-1}, X_n)\}$ is independent. By property CI11) $\{Y_{n+1}, U_{n-1}\}$ is conditionally independent, given X_n. Hence, we have for any n, any Borel set Q,

$$\begin{aligned} E[I_Q(X_{n+1}) | X_n, U_{n-1}] &= E\{I_Q[g_{n+1}(X_n, Y_{n+1})] | X_n, U_{n-1}\} \\ &= E\{I_Q[g_{n+1}(X_n, Y_{n+1})] | X_n\} && \text{by CI7)} \\ &= E[I_Q(X_{n+1}) | X_n] \end{aligned}$$

which establishes the Markov property. Now

$$\begin{aligned} P(X_{n+1} \in Q | X_n = u) &= E[I_Q(X_{n+1}) | X_n = u] \\ &= E\{I_Q[g_{n+1}(X_n, Y_{n+1})] | X_n = u\} \\ &= E\{I_Q[g_{n+1}(u, Y_{n+1})]\} && \text{by CE11)} \\ &= P[g_{n+1}(u, Y_{n+1}) \in Q] && \text{by E1a).} \end{aligned}$$ []

If $g_{n+1} = g$, invariant with n, and if $\{Y_n: 1 \leq n\}$ is independent, identically distributed, then $P(X_{n+1} \in Q | X_n = u) = P[g(u, Y_{n+1}) \in Q]$ is invariant with n. To illustrate, we consider the inventory problem above (c.f. Hillier and Lieberman [1974], Secs 8.17, 8.18).

Example E1-c (continued)

Suppose $m = 1$, $M = 3$, and D_n has the Poisson distribution with $\lambda = 1$. Then the state space $S = \{0, 1, 2, 3\}$ and $P(X_{n+1} = j | X_n = i) = P[g(i, D_{n+1}) = j]$. $g(0, D_{n+1}) = \max\{(3 - D_{n+1}), 0\}$.

Since $g(0, D_{n+1}) = 0$ iff $D_{n+1} \geq 3$,

$P(X_{n+1} = 0 | X_n = 0) = P(D_{n+1} \geq 3) = 0.0803$ (from table).

Since $g(0, D_{n+1}) = 1$ iff $D_{n+1} = 2$,

$P(X_{n+1} = 1 | X_n = 0) = P(D_{n+1} = 2) = 0.1839$ (from table).

Continuing in this way, we determine each transition probability and hence the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.0803 & 0.1839 & 0.3679 & 0.3679 \\ 0.6321 & 0.3679 & 0 & 0 \\ 0.2642 & 0.3679 & 0.3679 & 0 \\ 0.0803 & 0.1839 & 0.3679 & 0.3679 \end{bmatrix}.$$ []

The calculation procedure based on the equation $P(X_{n+1} = j | X_n = i) = P[g_{n+1}(i, Y_{n+1}) = j]$ can be justified in elementary terms for many special cases. The general result in Theorem E1-2 shows how the desired conditional independence of the past and future, given the present, arises out of the independence of the sequence $\{Y_n: 1 \leq n\}$ and establishes the validity of the calculation procedure in any situation (including continuous state space).

2. Markov chains with costs and rewards

In a variety of sequential decision making situations, the progression of states of the system in successive time periods can be represented in a useful way by a Markov process. The Markov character arises from the "memoryless" nature of the process. Often such sequential systems have a reward structure. Associated with each possible transition from one state to another is a "reward" (which may be negative). Consider the following classical example, utilized by Howard [1960] in his pioneering work in the area.

Example E2-a

The manufacturer of a certain item finds the market either "favorable" or "unfavorable" to his product in a given sales period. These conditions may be represented as state 0 or state 1, respectively. If the market is favorable in one period and is again favorable in the next period (transition from state 0 to state 0), the manufacturer's earnings are r_{00}. If the market is favorable in one period and unfavorable in the next (transition from state 0 to state 1), the earnings for the period are a smaller amount r_{01}. Similarly, the other possibilities have associated rewards. If the succession of states can be modeled by a Markov chain with stationary transition probabilities, then the system is characterized by two entities: the transition probability matrix **P** and the reward matrix **R**, given by

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} r_{00} & r_{01} \\ r_{10} & r_{11} \end{bmatrix}. \qquad []$$

We may express a general model for such a system as follows:

Let $\{X_n: 0 \leq n\}$ by a discrete-parameter Markov process with finite state

space S. The reward structure is expressed by the sequence $\{R_n: 1 \leq n\}$ of random variables

A) $R_{n+1} = r(X_n, X_{n+1})$, where $r(i,j) = r_{ij}$.

We are assuming that neither the reward structure nor the transition probabilities change with time. While more general situations could be modeled, we use the time-invariant case in subsequent developments.

Let $q_i = E[R_{n+1}|X_n = i]$ = expected reward in the next period, given the present state is i.

Then $q_i = E[r(X_n,X_{n+1})|X_n = i]$

$= E[r(i, X_{n+1})|X_n = i]$ by CE10)

$= \sum_j r(i,j)p_{ij}$.

Put $R_n^{(m)} = R_{n+1} + R_{n+2} + \ldots + R_{n+m}$ = total reward in the next m periods

Now

A1) $E[R_{n+k}|X_n = i] = E\{E[R_{n+k}|X_{n+1}]|X_n = i\}$ by CI8)

$= \sum_j E[R_{n+k}|X_{n+1} = j]p_{ij}$.

From this it follows that

$$E[R_n^{(m)}|X_n = i] = E[R_{n+1}|X_n = i] + \sum_j E[R_{n+1}^{(m-1)}|X_{n+1} = j]p_{ij}.$$

If we put

$v_i^{(m)} = E[R_n^{(m)}|X_n = i]$ (invariant with n in the stationary case)

we have

A2) $v_i^{(m)} = q_i + \sum_j p_{ij}v_j^{(m-1)}$, with $v_i^{(1)} = q_i$.

A second type of reward structure is exhibited in the following class of processes, which include inventory models of the type illustrated in Example E1-c.

Let $\{X_n: 0 \leq n\}$ be a constant Markov chain with finite state space, and let $\{D_{n+1}: 1 \leq n\}$ be an independent, identically distributed class such that for each $n \geq 0$, $\{D_{n+1},U_n\} = \{D_{n+1}, (X_0, X_1, \ldots, X_n)\}$ is an

independent pair. The associated reward structure is expressed by the process $\{R_n: 1 \leq n\}$, with

B) $R_{n+1} = r(X_n, D_{n+1})$.

Property CE11) shows that

$$q_i = E[R_{n+1} | X_n = i] = E[r(i, D_{n+1})] \quad \text{(invariant with } n\text{)}.$$

The hypothesis $\{D_{n+k}, U_{n+k-1}\}$ is independent and property CI11) ensure that $\{D_{n+k}, X_i\}$ is conditionally independent, given X_j, for $0 \leq i,j \leq n+k-1$, $i \neq j$. For fixed n, k, let

$$e(X_i) = E[R_{n+k} | X_i] = E[r(X_{n+k-1}, D_{n+k}) | X_i] \quad \text{for any } i \leq n + k - 1.$$

Then by CI8)

$$e(X_n) = E[e(X_i) | X_n] \text{ a.s. } n \leq i \leq n + k - 1.$$

Hence,

$$\text{B1)} \quad E[R_{n+k} | X_n = i] = E\{E[R_{n+k} | X_{n+1}] | X_n = i\}$$
$$= \sum_j E[R_{n+k} | X_{n+1} = j] p_{ij}.$$

Applying this formula for $k = 2, 3, \ldots, m$, we obtain

$$\text{B2)} \quad v_i^{(m)} = q_i + \sum_j p_{ij} v_j^{(m-1)} \quad \text{with} \quad v_i^{(1)} = q_i.$$

The identity of form of A1), B1) and A2), B2) shows that the following analysis holds for either type of reward structure.

Consider the average expected reward per period for m periods.

$$E[\tfrac{1}{m} R_n^{(m)}] = \frac{1}{m} \sum_{i=1}^{m} E[R_{n+i}]$$
$$= \frac{1}{m} \sum_{i=1}^{m} E(E\{E[R_{n+i} | X_{n+i-1}] | X_{n-1}\}) \quad \text{by CE1b) and CI8).}$$

Now $E[R_{n+i} | X_{n+i-1} = j] = q_j$, so that

$$E\{E[R_{n+i} | X_{n+i-1}] | X_{n-1} = k\} = \sum_j p_{kj}^{(i)} q_j,$$

where $p_{kj}^{(i)}$ is the i-step transition probability from k to j. Hence,

$$E[\tfrac{1}{m} R_n^{(m)}] = \frac{1}{m} \sum_{i=1}^{m} \sum_k [P(X_{n-1} = k) \sum_j p_{kj}^{(i)} q_j] = \sum_k [P(X_{n-1} = k) \sum_j q_j (\frac{1}{m} \sum_{i=1}^{m} p_{kj}^{(i)})].$$

If the Markov chain is constant, irreducible, aperiodic, as is usually the case, it is known that

$$\frac{1}{m}\sum_{i=1}^{m} p_{kj}^{(i)} \to \pi_j \quad \text{as} \quad m \to \infty \quad \text{(invariant in } k\text{)}.$$

Here π_j is the long-run probability that the process is in state j. Since the limit is invariant with k, we may sum out the $P(X_{n-1} = k)$ to obtain

3) $$\lim_{m\to\infty} E[\frac{1}{m} R_n^{(m)}] = \sum_j q_j\pi_j = g.$$

A similar argument shows that for each state i

4) $$\lim_{m\to\infty} E[\frac{1}{m} R_n^{(m)} | X_n = i] = \lim_{m\to\infty} \frac{1}{m} v_i^{(m)} = \sum_j q_j\pi_j = g.$$

Here g is the average gain or reward per period, in the long run. We illustrate by considering numerical values in the introductory examples.

Example E2-a (continued)

Suppose $\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 \\ 2/5 & 3/5 \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix}$ and $\mathbf{R} = \begin{bmatrix} 9 & 3 \\ 3 & -7 \end{bmatrix}$.

To find the long-run distribution, we solve the set of equations

$$5\pi_0 + 4\pi_1 = 10\pi_0$$
$$5\pi_0 + 6\pi_1 = 10\pi_1 \quad \text{to obtain the values } \pi_0 = 4/9 \text{ and } \pi_1 = 5/9$$
$$\pi_0 + \pi_1 = 1.$$

Then

$$q_0 = \sum_j p_{0j}r_{0j} = \frac{1}{2}9 + \frac{1}{2}3 = 6 \qquad q_1 = \sum_j p_{1j}r_{1j} = \frac{2}{5}3 + \frac{3}{5}(-7) = -3$$
$$g = \lim \frac{1}{m} v_i^{(m)} = \sum_j q_j\pi_j = 6\,\frac{4}{9} - 3\,\frac{5}{9} = 1.$$

[]

Example E1-c (continued)

Suppose $m = 1$ and $M = 3$, as before.

If k units are ordered, the cost is $10 + 25k$, $0 < k \leq M$.

If $k = 0$, the cost of ordering is zero.

For each unit of unsatisfied demand, a penalty of \$50 is assessed

We suppose the demand D_n in period n has Poisson distribution, with $\lambda = 1$. We may then calculate the cost function (negative of reward)

$$C(X_n, D_{n+1}) = \begin{cases} 10 + 25(M - X_n) + 50 \max\{(D_{n+1} - M), 0\} & \text{for } 0 \leq X_n < m \\ 50 \max\{(D_{n+1} - X_n), 0\} & \text{for } m \leq X_n \leq M. \end{cases}$$

Thus,

$$C(0, D_{n+1}) = 85 + 50 \max\{(D_{n+1} - 3), 0\}$$

$$C(i, D_{n+1}) = 50 \max\{(D_{n+1} - i), 0\} \quad \text{for } i = 1, 2, 3.$$

Now

$$q_0 = E[C(0, D_{n+1})] = 85 + 50\, E[I_{\{D \geq 3\}}(D - 3)]$$

$$= 85 + 50 \sum_{k=4}^{\infty} (k - 3)p_k \qquad \text{(term for } k = 3 \text{ is zero).}$$

For the Poisson distribution $\sum_{k=n}^{\infty} kp_k = \lambda \sum_{k=n-1}^{\infty} p_k$. Hence,

$$q_0 = 85 + 50[\sum_{k=3}^{\infty} p_k - 3 \sum_{k=4}^{\infty} p_k] = 86.2 \quad \text{(Using table for Poisson distribution).}$$

$$q_1 = E[C(1, D_{n+1})] = 50 \sum_{k=2}^{\infty} (k - 1)p_k = 50[\sum_{k=1}^{\infty} p_k - \sum_{k=2}^{\infty} p_k] = 50p_1 = 18.4.$$

Similarly, we obtain

$$q_2 = E[C(2, D_{n+1})] = 50 \sum_{k=3}^{\infty} (k - 2)p_k = 5.2$$

and

$$q_3 = E[C(3,D_{n+1})] = 50 \sum_{k=4}^{\infty} (k - 3)p_k = 1.2 .$$

To obtain the long-run probabilities, we utilize the fact that the convergence is rapid and consider $P^2, P^4, \ldots$ until results stabilize. Direct calculations of matrix products shows that

$$P^8 = \begin{bmatrix} 0.286 & 0.285 & 0.264 & 0.166 \\ 0.286 & 0.285 & 0.264 & 0.166 \\ 0.286 & 0.285 & 0.264 & 0.166 \\ 0.286 & 0.285 & 0.264 & 0.166 \end{bmatrix}$$

from which we conclude $\pi_0 = 0.286$, $\pi_1 = 0.285$, $\pi_2 = 0.264$, and $\pi_3 = 0.166$. These add to 1.001, indicating a small roundoff error. Utilizing these values, we obtain

$$g = \lim_{m\to\infty} \frac{1}{m} v_i^{(m)} = \sum_j q_j \pi_j = 31.5 . \qquad []$$

The treatment, once equations A1), A2) or B1), B2) and 3), 4) are obtained, is standard. As a matter of fact, we have used examples taken from published texts. In most standard works, the derivations are intuitive and incomplete. We have provided a development based on fundamental assumptions of independence and conditional independence (or Markov conditions). Such a development should both sharpen intuition and provide a sound mathematical basis for utilizing the models.

3. Continuous-parameter Markov processes

There are certain technical difficulties in the theory of continuous-parameter processes. However, advanced methods show that a process can be determined essentially for applications if all finite-dimensional distributions are determined (i.e., if the joint distribution for any finite subclass of the random variables is determined).

Consider a real process $\{X_t: t \geq 0\}$ (i.e., $T = [0, \infty)$). Let U, V, W be finite subsets of T: $U = \{u_1, u_2, \ldots, u_m\}$, $V = \{v_1, v_2, \ldots, v_n\}$, and $W = \{w_1, w_2, \ldots, w_q\}$. We suppose $u_i < u_{i+1}$, $v_j < v_{j+1}$, and $w_k < w_{k+1}$ for all indicated i, j, k. We say U precedes V, denoted $U \prec V$, iff every element of U is less than every element of V. We put $X_U = (X_{u_1}, X_{u_2}, \ldots, X_{u_m})$, $X_V = (X_{v_1}, X_{v_2}, \ldots, X_{v_n})$, and $X_W = (X_{w_1}, X_{w_2}, \ldots, X_{w_q})$.

DEFINITION. The process $\{X_t: t \geq 0\}$ is a Markov process iff for any $U \prec \{v\} \prec \{w\}$ we have

M) $E[I_M(X_w)|X_v, X_U] = E[I_M(X_w)|X_v]$ a.s. for all Borel Sets M on the codomain of X_w (i.e., in the state space S).

It is clear that condition M) is equivalent to

M') For any finite $U \prec \{v\} \prec \{w\}$, $\{X_w, X_U\}$ is conditionally independent, given X_v.

As in the discrete-parameter case, we have the equivalent condition (see Theorem E1-1)

M'') For any finite $U \prec V \prec W$ in T, $\{X_W, X_U\}$ is conditionally independent, given X_V.

These and other equivalent expressions for the conditional independence condition provide major tools for the study of Markov processes.

Many of the Markov processes encountered in practice may be recognized by virtue of the following property.

DEFINITION. A random process $\{X_t: t \in T\}$ has <u>independent increments</u> iff for each finite subset $T_n = \{t_0, t_1, \ldots, t_n\}$ of the parameter set T, with $t_0 < t_1 < \ldots < t_n$, the class $\{X_{t_0}, X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}\}$ of random variables is independent.

Two of the most widely studied and utilized random processes have this property.

<u>Poisson process</u>.

The parameter set is $T = [0, \infty)$. The process counts the number of occurrences of some phenomenon in given time intervals. The random variable X_t counts the number of occurrences in time interval $(0,t]$. We set $X_0 = 0$. Then $X_t - X_s$, for $s < t$, is the number of occurrences in the time interval $(s,t]$. The property of independent increments models the fact that the numbers of occurrences in nonoverlapping time intervals are independent. What happens in one interval is not affected by and has no effect on what happens in other intervals.

<u>Wiener process</u> (Brownian motion).

The parameter set is $T = [0, \infty)$. $X_0 = 0$. The process is a model of the movement along a line of a "particle" under "random disturbances." X_t is the net movement along a coordinate axis in the time interval $(0, t]$. In many situations, the disturbances are of such a character that the distances moved in disjoint time intervals may be assumed independent. Hence, the independent-increment assumption is appropriate.

In the discrete-parameter case, the class of random walks (see Example E1-a) possess the independent-increment property. We have $X_n = Y_1 + Y_2 + \ldots + Y_n$ and $X_{m+k} - X_m = Y_{m+1} + Y_{m+2} + \ldots + Y_{m+k}$. The assumed independence of the class $\{Y_i : 1 \leq i\}$ ensures independence of the increments.

We wish to show that a process with independent increments is a Markov process. To facilitate exposition, we adopt the following terminology and notation.

1) We say $T_n = \{t_0, t_1, \ldots, t_n\} \subset T$ is a strictly ordered, finite subset of T iff $t_0 < t_1 < \ldots < t_n$.

2) For any strictly ordered, finite subset of T, we define the random variables $Y_0 = X_{t_0}$ and $Y_k = X_{t_k} - X_{t_{k-1}}$ for $1 \leq k \leq n$, and the random vectors $U_k = (X_{t_0}, X_{t_1}, \ldots, X_{t_k})$ and $Z_k = (Y_0, Y_1, \ldots, Y_k)$ for each k, $1 \leq k \leq n$.

We note that if we have the values of the coordinates of any one of the vectors U_n, Z_n, (Z_{n-1}, X_{t_n}), or (U_{n-1}, Y_n) the values of the coordinates of the others are obtained by linear transformations, which are continuous, hence Borel. Thus, we may assert

A) Any one of the random vectors U_n, Z_n, (Z_{n-1}, X_{t_n}), or (U_{n-1}, Y_n) is a Borel function of any one of the others.

By virtue of property CE9b), we have

B) $E[W|Z_n] = E[W|U_n] = E[W|U_{n-1}, X_{t_n}] = E[W|Z_{n-1}, X_{t_n}] = E[W|U_{n-1}, Y_n]$ a.s.

Also, by virtue of independence of Borel functions of independent random vectors,

C) If any of the pairs $\{Y_{n+1}, U_n\}$, $\{Y_{n+1}, Z_n\}$, $\{Y_{n+1}, (Z_{n-1}, X_{t_n})\}$, is independent, so are the others.

With these facts, we can now establish the fundamental result

Theorem E3-1

If the process $\{X_t: t \in T\}$ has independent increments, then it is a Markov process.

PROOF

We show that for any strictly ordered, finite $T_n \subset T$, the condition M') holds for $X_U = U_{n-1}$, $X_V = X_{t_n}$, and $X_W = X_{t_{n+1}}$.

Now $g(X_{t_{n+1}}) = g(X_{t_n} + Y_{n+1}) = h(X_{t_n}, Y_{n+1})$, with h Borel and

$$E[g(X_{t_{n+1}})|U_{n-1}, X_{t_n}] = E[h(X_{t_n}, Y_{n+1})|Z_{n-1}, X_{t_n}] \text{ a.s. by proposition B)}$$

By proposition C) and CI11), $\{Y_{n+1}, Z_{n-1}\}$ is conditionally independent, given X_{t_n}. Hence,

$$E[h(X_{t_n}, Y_{n+1})|Z_{n-1}, X_{t_n}] = E[h(X_{t_n}, Y_{n+1})|X_{t_n}] \text{ a.s. by CI7)}$$

We may therefore assert

$$E[g(X_{t_{n+1}})|U_{n-1}, X_{t_n}] = E[g(X_{t_{n+1}})|X_{t_n}] \text{ a.s.}$$

which is the desired property. []

The following alternate criterion for independent increments is frequently useful as an assumption in modeling.

Theorem E3-2

A process $\{X_t: t \in T\}$ has independent increments iff for every strictly ordered, finite $T_n \subset T$, the pair $\{Y_n, U_{n-1}\}$ is independent.

PROOF

a) If the process has independent increments, the pair $\{Y_n, Z_{n-1}\}$ is independent. By proposition C), above, so is $\{Y_n, U_{n-1}\}$ an independent pair.

b) Suppose $\{Y_n, U_{n-1}\}$ is independent for all T_n. Let T_n be arbitrarily selected, but fixed. For each k, $0 \leq k \leq n$, set $T_k = \{t_0, t_1, \ldots, t_k\}$. By hypothesis, $\{Y_k, U_{k-1}\}$ is independent, $1 \leq k \leq n$. By proposition C), the pair $\{Y_k, Z_{k-1}\}$ is independent, $1 \leq k \leq n$. In particular, $\{Y_1, Z_0\} = \{Y_1, Y_0\}$ is independent.

Suppose for some $k \geq 2$, $\{Y_0, Y_1, \ldots, Y_{k-1}\}$ is independent. Then by the independence of $\{Y_k, Z_{k-1}\}_{k-1} = \{Y_k, (Y_0, Y_1, \ldots, Y_{k-1})\}$, we have $P(\bigcap_{i=0}^{k} Y_i \in M_i) = P(Y_k \in M_k)P(\bigcap_{i=0}^{k-1} Y_i \in M_i) = P(Y_k \in M_k) \prod_{i=1}^{k-1} P(Y_i \in M_i)$. Thus, $\{Y_0, Y_1, \ldots, Y_k\}$ is independent. By mathematical induction, the class $\{Y_0, Y_1, \ldots, Y_n\}$ is independent. Since T_n is arbitrary, the desired proposition follows. []

4. The Chapman-Kolmogorov equation

For a Markov process $\{X_t: 0 \le t\}$, let $0 \le s < t < u$. Then the pair $\{X_s, X_u\}$ is conditionally independent, given X_t. As a special case of CI8), we have

CK) $E[g(X_u)|X_s] = E\{E[g(X_u)|X_t]|X_s\}$ a.s.

This is the Chapman-Kolmogorov equation, which plays a significant role in the study of Markov processes.

For a chain with finite state space S, the equation takes a simple form which is usually determined from the first form of the Markov property in Sec E1 and elementary probability patterns. If $p_{jk}(s,t) = P(X_t = k|X_s = j)$, the Chapman-Kolmogorov equation is usually written

CK') $p_{ik}(s,u) = \sum_j p_{ij}(s,t)p_{jk}(t,u) \qquad 0 \le s < t < u.$

To see that this is a special form of CK), note that

$$\begin{aligned} P(X_u = k|X_s = i) &= E[I_{\{k\}}(X_u)|X_s = i] \\ &= E\{E[I_{\{k\}}(X_u)|X_t]|X_s = i\} \\ &= \sum_j E[I_{\{k\}}(X_u)|X_t = j]P(X_t = j|X_s = i) \\ &= \sum_j p_{ij}(s,t)p_{jk}(t,u). \end{aligned}$$

In the case of stationary transition probabilities, let $p_{ik}^{(m)}$ be the m-step transition probability from state i to state k. CK') becomes

CK") $p_{ik}^{(m+n)} = \sum_j p_{ij}^{(m)} p_{jk}^{(n)}$

which is the form commonly encountered in elementary treatments. In such treatments, the transition probability matrix $\mathbf{P}$ plays a central role. If $\mathbf{P}^{(m)}$ is the matrix of m-step transition probabilities, then $\mathbf{P}^{(m)} = \mathbf{P}^m = \mathbf{PPP}...\mathbf{P}$ (m factors). The Chapman-Kolmogorov equation CK") may be expressed compactly as

CK") $\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)}\mathbf{P}^{(n)}$.

If the random variables are absolutely continuous, the Chapman-Kolmogorov equation is often expressed in terms of conditional density functions.

CK''') $f_{X_u|X_s}(z|x) = \int f_{X_u|X_t}(z|y) f_{X_t|X_s}(y|x)\, dy$.

In this case CK) may be written

$$\int g(z) f_{X_u|X_s}(z|x)\, dz = \int [\int g(z) f_{X_u|X_t}(z|y)\, dz] f_{X_t|X_s}(y|x)\, dy$$
$$= \int g(z) [\int f_{X_u|X_t}(z|y) f_{X_t|X_s}(y|x)\, dy]\, dz.$$

In order for this equation to hold for all Borel functions g, by an analog to property E7) for integrals on the real line, we must have CK''') for each x.

In spite of the importance of the Chapman-Kolmogorov equation in many aspects of Markov process theory, it is <u>not</u> true that the validity of this equation implies the process is Markov. Stated another way, it is not true that the condition CI7) may be replaced by the condition $E[g(X)|Z,Y] = E[g(X)|Z]$ a.s. for any Borel function g. The latter condition is not sufficient for the conditional independence of {X,Y}, given Z. W. Feller has given counterexamples. The following is taken from Parzen [1962], p 203, but it is due essentially to Feller.

<u>Example</u> E4-a

Consider a sequence of containers, each with four balls, numbered one through four. Select a ball independently, on an equally likely basis, from each container. Let

$A_m(1)$ = event ball 1 or 4 is drawn from the <u>mth</u> container

$A_m(2)$ = event ball 2 or 4 is drawn from the <u>mth</u> container

$A_m(3)$ = event ball 3 or 4 is drawn from the <u>mth</u> container.

Under the usual assumptions, $P[A_m(j)] = 1/2$ for any $m \geq 1$, any $j = 1, 2,$ or 3. For any m (i.e., any container), we have a classical

example of a class $\{A_m(j): j = 1, 2, 3\}$ of events which is pairwise independent, but not independent. Since selections from various containers are independent, we assume $\{A_m(j_m): 1 \leq m\}$ is an independent class for any sequence $\{j_m: 1 \leq m\}$ of elements of the set $\{1, 2, 3\}$. Thus we may assert that $\{A_m(j): 1 \leq m, j = 1, 2, 3\}$ is a pairwise independent class, with $P[A_m(j)] = 1/2$ for any permissible m, j. We now form the process $\{X_n: 1 \leq n\}$ by setting

$$X_{3(m-1)+j} = I_{A_m(j)}, \quad j = 1, 2, 3, \quad m \geq 1.$$

This process has state space $S = \{0, 1\}$, and the members are pairwise independent, with $P(X_n = 0) = P(X_n = 1) = 1/2$. We also have

$P(X_{n+r} = j | X_n = i) = P(X_{n+r} = j) = 1/2$ for any $j, k \in \{0, 1\}$, any $n \geq 1$, any $r \geq 1$.

Thus, the m-step transition probability matrix is

$$\mathbf{P}^{(m)} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{for any } m \geq 1.$$

Easy matrix calculations show

$$\mathbf{P}^{(m)}\mathbf{P}^{(n)} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \mathbf{P}^{(m+n)}$$

so the Chapman-Kolmogorov equation holds. However, the process is not Markov, as the following argument shows. Since $A_{m+1}(1)A_{m+1}(2)$ is a subset of $A_{m+1}(3)$, we have

$$P(X_{3m+3} = 1 | X_{3m+2} = 1, X_{3m+1} = 1) = P(A_{m+1}(3) | A_{m+1}(2)A_{m+1}(1)) = 1$$
$$\neq P(X_{3m+3} = 1 | X_{3m+2} = 1) = 1/2. \quad []$$

5. Proof of a basic theorem on Markov processes

We utilize the notational scheme introduced in Sec E1. To prove Theorem E1-1, we first obtain an intermediate result.

Theorem E5-1

For a Markov process $\{X_t: t \in T\}$, with $T = \{0, 1, 2, \ldots\}$, the pair $\{X_{t+1}, U_{s-1}\}$ is conditionally independent, given any $V_{s,t}$, $1 \leq s \leq t$.

PROOF

We note that $U_{t-1} = (U_{s-1}, V_{s,t-1})$ and $V_{s,t} = (V_{s,t-1}, X_t)$. For any Borel function g, any s, t, $1 \leq s \leq t$,

$$E[g(X_{t+1})|X_t] = E[g(X_{t+1})|U_{t-1}, X_t] \quad \text{a.s.} \qquad \text{by M') and CI6)}$$
$$= E[g(X_{t+1})|U_{s-1}, V_{s,t}].$$

By Lemma D5-1, with $V = X_t$, $U = U_{t-1}$, $Z = V_{s,t-1} = h(U_{t-1})$,

$$E[g(X_{t+1})|X_t] = E[g(X_{t+1})|V_{s,t-1}, X_t] \quad \text{a.s.}$$
$$= E[g(X_{t+1})|V_{s,t}].$$

The theorem follows by CI6). []

Theorem E1-1

A process $\{X_t: t \in T\}$, $T = \{0, 1, 2, \ldots\}$, is Markov iff

M'') $\{W^*_{t+1,t+n}, U^*_{s-1}\}$ is conditionally independent, given any finite extended present $V_{s,t}$, $1 \leq s \leq t$, any $n \geq 1$, any $W^*_{t+1,t+n}$, U^*_{s-1}.

PROOF

M'') implies M') as a special case.

Suppose M') holds. We need only establish M*) $\{W_{t+1,t+n}, U_{s-1}\}$ is conditionally independent, given $V_{s,t}$, $1 \leq s \leq t$, any $n \geq 1$. The more general condition follows from CI9), with $W^*_{t+1,t+n} = h(W_{t+1,t+n})$ and $U^*_{s-1} = k(U_{s-1})$. We construct a proof by mathematical induction on n, utilizing Theorem E5-1.

i) Since $X_{t+1} = W_{t+1,t+1}$, M*) holds for $n = 1$, by Theorem E5-1.

ii) Suppose M*) holds for $n = k$.

By Theorem E5-1, $\{X_{t+k+1}, U_t\}$ is conditionally independent, given $W_{t+1,t+k}$. Hence, for any Borel function g,

$$
\begin{aligned}
&E[g(W_{t+1,t+k+1})|U_{s-1},V_{s,t}] \\
&\quad = E[g(W_{t+1,t+k}, X_{t+k+1})|U_t] \\
&\quad = E\{E[g(W_{t+1,t+k}, X_{t+k+1})|W_{t+1,t+k}]|U_t\} \quad \text{by CI8)} \\
&\quad = E[e(W_{t+1,t+k})|U_{s-1},V_{s,t}] \\
&\quad = E[e(W_{t+1,t+k})|V_{s,t}] \quad \text{by inductive hypothesis and CI6)} \\
&\quad = E\{E[g(W_{t+1,t+k}, X_{t+k+1})|W_{t+1,t+k}]|V_{s,t}\} \\
&\quad = E[g(W_{t+1,t+k+1})|V_{s,t}] \text{ a.s.} \quad \text{by CI8) and CI9).}
\end{aligned}
$$

By CI6), M*) holds for $n = k + 1$.

iii) By mathematical induction, M*) holds for any $n \geq 1$. []

6. Problems

E-1 Stopping times. In dealing with a random process $\{X_n: 0 \le n\}$ it is sometimes desirable to consider a randomly selected member of the process. Suppose, for example, we wish to stop the process when a certain result (or pattern of results) is observed. This means we select X_n as the last variable iff the observed sequence $(s_0, s_1, \ldots, s_n) \in S^{n+1}$ of results exhibits a prescribed pattern, hence belongs to a certain subset M_n of S^{n+1}. We use this to formalize the notion as follows:

DEFINITION. A nonnegative, integer-valued random variable T is called a stopping time for the process $\{X_n: 0 \le n\}$ iff the event $A_k = \{\omega: T(\omega) = k\}$ is determined by $U_k = (X_0, X_1, \ldots, X_k)$. Thus, $A_k = U_k^{-1}(M_k)$, with $A_k A_j = \emptyset$ for $k \ne j$. We assume $\sum_{k=0}^{\infty} P(A_k) = 1$, which means that with probability one T is finite.

It is apparent that $T = \sum_{k=0}^{\infty} k\, I_{A_k} = \sum_{k=0}^{\infty} k\, I_{M_k}(U_k)$ a.s.

a) Suppose X_n is the value of a critical dimension of the nth item from a production line. The desired value is a. The process is stopped for readjustment whenever $|X_n - a| > b$. Show that if T is the random variable which designates the number of the item at which the line is stopped, then T is a stopping time for the process.

Suggestion. Express M_k in terms of the coordinate sets $M = [a-b, a+b]$.

b) Show that if the X_n are integer-valued, the random variable T_1 defined by $T_1(\omega) = \min\{n > 0: X_n(\omega) = i\}$ is a stopping time.

c) Show that if T_1 is a stopping time for an integer-valued process, so is T_2 defined by $T_2(\omega) = \min\{n > T_1(\omega): X_n(\omega) = i\}$.

E-2 Suppose T is a stopping time for the process $\{X_n: 0 \le n\}$. Let $U_T = \sum_{k=0}^{\infty} U_k I_{A_k} = \sum_{k=0}^{\infty} U_k I_{M_k}(U_k)$. The expressions $g(U_T)$ and $I_Q(U_T)$ must be interpreted, since the dimension of random vector U_T changes with T. If $Q \in S^{\infty}$ and $Q(k)$ is the projection onto S^{k+1}, then we set $I_Q(U_T) = \sum_k I_{Q(k)}(U_k) I_{M_k}(U_k)$. Similarly $g(U_T) = \sum_k g_k(U_k) I_{M_k}(U_k)$. Show that $E[g(Y)|U_T] = \sum_k E[g(Y)|U_k] I_{M_k}(U_k)$ a.s.

E-3 Strong Markov property. Suppose $\{X_n: 0 \le n\}$ is a Markov process and T is a stopping time for the process.

a) Show that $E[g(W_{T,T+n})|U_T] = E[g(W_{T,T+n})|X_T]$
$= \sum_k E[g(W_{k,k+n})|X_k] I_{M_k}(U_k)$ a.s.

b) If the process is homogeneous, show that

$$E[g(W_{T,T+n})|X_T] = E[g(W_{0,n})|X_0] \quad \text{a.s.}$$

E-4 Martingales. The following class of random processes has many connections with the class of Markov processes (cf Karlin and Taylor [1975], Chap 6).

DEFINITION. Let $\{X_n: 0 \le n\}$ be a sequence of real random variables and $\{Y_n: 0 \le n\}$ be a sequence of random vectors. Then $\{X_n: 0 \le n\}$ is a martingale with respect to $\{Y_n: 0 \le n\}$ iff i) $E[|X_n|]$ is finite for each $n \ge 0$, and ii) $E[X_{n+1}|Y_0, Y_1, \ldots, Y_n] = X_n$ a.s. for each $n \ge 0$.

Note that conditions i) and ii) imply iii) $X_n = e_n(Y_0, Y_1, \ldots, Y_n)$ a.s., with e_n a Borel function for any $n \ge 0$. If $Y_k = X_k$, all k, we say $\{X_n: 0 \le n\}$ is a martingale, without qualifying expression.

a) Show that for a martingale $E[X_n] = E[X_0]$ for all n.

b) Show that if $\{X_n: 0 \le n\}$ has independent increments (hence is Markov) and $E[X_n] = E[X_0]$ all $n \ge 0$, the process is a martingale.

Appendices

APPENDICES

APPENDIX I. Properties of Mathematical Expectation

E1) $E[I_A] = P(A)$.

E1a) $E[I_M(X)] = P(X \in M)$; $E[I_M(X)I_N(Y)] = P(X \in M, Y \in N)$ (with extension by mathematical induction to any finite number of factors).

E2) Linearity. $E[aX + bY] = aE[X] + bE[Y]$ (with extension by mathematical induction to any finite linear combination).

E3) Positivity; monotonicity.

a) $X \geq 0$ a.s. implies $E[X] \geq 0$, with equality iff $X = 0$ a.s.

b) $X \geq Y$ a.s. implies $E[X] \geq E[Y]$, with equality iff $X = Y$ a.s.

E4) Monotone convergence. If $X_n \to X$ monotonically a.s., then $E[X_n] \to E[X]$ monotonically.

E5) Independence. The pair $\{X,Y\}$ of random vectors is independent iff $E[I_M(X)I_N(Y)] = E[I_M(X)]E[I_N(Y)]$ for all Borel sets M, N on the codomains of X, Y, respectively,

iff $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for all real-valued Borel functions g, h such that the expectations exist.

E6) Uniqueness.

a) Suppose Y is a random vector with codomain R^m and g, h are real-valued Borel functions on the range of Y. If $E[I_M(Y)g(Y)] = E[I_M(Y)h(Y)]$ for all Borel sets M on the codomain of Y, then $g(Y) = h(Y)$ a.s.

b) More generally, if $E[I_M(Y)I_N(Z)g(Y,Z)] = E[I_M(Y)I_N(Z)h(Y,Z)]$ for all Borel sets M, N in the codomains of Y, Z, respectively, then $g(Y,Z) = h(Y,Z)$ a.s.

E7) Fatou's lemma. If $X_n \geq 0$ a.s., then $E[\liminf X_n] \leq \liminf E[X_n]$.

E8) Dominated convergence. If $X_n \to X$ a.s. and $|X_n| \leq Y$ a.s., for each n, with $E[Y]$ finite, then $E[X_n] \to E[X]$.

9) Countable additivity. Suppose $E[X]$ exists and $A = \biguplus_{i=1}^{\infty} A_i$. Then $E[I_A X] = \sum_{i=1}^{\infty} E[I_{A_i} X]$.

10) Existence. If $E[g(X)]$ is finite, then there is a real-valued Borel function e, unique a.s. $[P_Y]$, such that $E[I_M(Y)g(X)] = E[I_M(Y)e(Y)]$ for all Borel sets M in the codomain of Y.

11) Triangle inequality. $|E[g(X)]| \leq E[|g(X)|]$.

12) Mean-value theorem. If $a \leq X \leq b$ a.s. on A, then $aP(A) \leq E[I_A X] \leq bP(A)$.

13) Let g be a nonnegative Borel function, defined on the range of X. Let $A = \{\omega: g[X(\omega)] \geq a\}$. Then $E[g(X)] \geq aP(A)$.

14) Markov's inequality. If $g \geq 0$ and nondecreasing for $t \geq 0$ and $a \geq 0$, then $g(a)P(|X| \geq a) \leq E[g(|X|)]$.

15) Jensen's inequality. If g is a convex function on an interval I which includes the range of real random variable X, then $g(E[X]) \leq E[g(X)]$.

16) Schwarz' inequality. If X, Y are real or complex random variables with $E[|X|^2]$ and $E[|Y|^2]$ finite, then $|E[XY]|^2 \leq E[|X|^2]E[|Y|^2]$, with equality iff there is a constant c such that $X = cY$ a.s.

17) Hölder's inequality. Let $1 \leq p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. If X, Y are real or complex random variables with $E[|X|^p]$ and $E[|Y|^q]$ finite, then $E[|XY|] \leq E[|X|^p]^{1/p} E[|Y|^q]^{1/q}$.

18) Minkowski's inequality. Let $1 \leq p < \infty$. If X, Y are real or complex random variables with $E[|X|^p]$ and $E[|Y|^p]$ finite, then $E[|X + Y|^p]^{1/p} \leq E[|X|^p]^{1/p} + E[|Y|^p]^{1/p}$.

APPENDIX II. Properties of Conditional Expectation, given a Random Vector

We suppose, without repeated assertion, that the random vectors and Borel functions in the expressions below are such that ordinary expectations exist.

CE1) $e(Y) = E[g(X)|Y]$ a.s. iff $E[I_M(Y)g(X)] = E[I_M(Y)e(Y)]$ for all Borel sets M on the codomain of Y.

CE1a) If $P(Y \in M) > 0$, then $E[I_M(Y)e(Y)] = E[g(X)|Y \in M]P(Y \in M)$.

CE1b) $E[g(X)] = E\{E[g(X)|Y]\}$.

CE2) Linearity. $E[ag(X) + bh(Y)|Z] = aE[g(X)|Z] + bE[h(Y)|Z]$ a.s. (with extension by mathematical induction to any finite linear combination).

CE3) Positivity; monotonicity.

$g(X) \geq 0$ a.s. implies $E[g(X)|Y] \geq 0$ a.s.

$g(X) \geq h(Y)$ a.s. implies $E[g(X)|Z] \geq E[h(Y)|Z]$ a.s.

CE4) Monotone convergence. $X_n \to X$ a.s. monotonically implies $E[X_n|Y] \to E[X|Y]$ a.s. monotonically.

CE5) Independence. a) $\{X,Y\}$ is an independent pair iff

b) $E[I_N(X)|Y] = E[I_N(X)]$ a.s. for all Borel sets N iff

c) $E[g(X)|Y] = E[g(X)]$ a.s. for all Borel functions g.

CE6) $e(Y) = E[g(X)|Y]$ a.s. iff $E[h(Y)g(X)] = E[h(Y)e(Y)]$ for all Borel h.

CE7) If $X = h(Y)$, then $E[g(X)|Y] = g(X)$ a.s. for all Borel g.

CE8) $E[h(Y)g(X)|Y] = h(Y)E[g(X)|Y]$ a.s.

CE9) If $Y = h(W)$, then $E\{E[g(X)|Y]|W\} = E\{E[g(X)|W]|Y\} = E[g(X)|Y]$ a.s.

CE9a) $E\{E[g(X)|Y]|Y,Z\} = E\{E[g(X)|Y,Z]|Y\} = E[g(X)|Y]$ a.s.

CE9b) If $Y = h(W)$, where h is Borel with a Borel inverse, then $E[g(X)|Y] = E[g(X)|W]$ a.s.

CE10) If g is Borel such that $E[g(X,v)]$ is finite for all v on the the range of Y and $E[g(X,Y)]$ is finite, then

$E[g(X,Y)|Y = u] = E[g(X,u)|Y = u]$ a.s. $[P_Y]$.

CE11) In CE10), if $\{X,Y\}$ is an independent pair, then

$E[g(X,Y)|Y = u] = E[g(X,u)]$ a.s. $[P_Y]$.

CE12) Triangle inequality. $|E[g(X)|Y]| \leq E[|g(X)||Y]$ a.s.

CE13) Jensen's inequality. If g is a convex function on an interval I which contains the range of real random variable X, then

$g(E[X|Y]) \leq E[g(X)|Y]$ a.s.

APPENDIX III. Properties of Conditional Independence, given a Random Vector

The following conditions are equivalent:

CI1) $E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z]E[I_N(Y)|Z]$ a.s. $\forall$ Borel sets M, N.

CI2) $E[I_M(X)|Z,Y] = E[I_M(X)|Z]$ a.s. $\forall$ Borel sets M.

CI3) $E[I_M(X)I_Q(Z)|Z,Y] = E[I_M(X)I_Q(Z)|Z]$ a.s. $\forall$ Borel sets M, Q.

CI4) $E[I_M(X)I_Q(Z)|Y] = E\{E[I_M(X)I_Q(Z)|Z]|Y\}$ a.s. $\forall$ Borel sets M, Q.

CI5) $E[g(X)h(Y)|Z] = E[g(X)|Z]E[h(Y)|Z]$ a.s. $\forall$ Borel functions g, h.

CI6) $E[g(X)|Z,Y] = E[g(X)|Z]$ a.s. $\forall$ Borel functions g.

CI7) $E[g(X,Z)|Z,Y] = E[g(X,Z)|Z]$ a.s. $\forall$ Borel functions g.

CI8) $E[g(X,Z)|Y] = E\{E[g(X,Z)|Z]|Y\}$ a.s. $\forall$ Borel functions g.

DEFINITION. The pair of random vectors {X,Y} is conditionally independent, given Z, iff the product rule CI1) holds. An arbitrary class of random vectors is conditionally independent, given Z, if an analogous product rule holds for each finite subclass of two or more members of the class.

CI9) If {X,Y} is conditionally independent, given Z, U = h(X), and V = k(Y), with h, k Borel, then {U,V} is conditionally independent, given Z.

CI10) If the pair {X,Y} is conditionally independent, given Z, then

a) $E[g(X)h(Y)] = E\{E[g(X)|Z]E[h(Y)|Z]\} = E[e_1(Z)e_2(Z)]$

b) $E[g(X)|Y \in N]P(Y \in N) = E\{E[I_N(Y)|Z]E[g(X)|Z]\}$.

CI11) If {Y, (X,Z)} is independent, then {X,Y} is conditionally independent, given Z.

CI12) If {X,Y} is conditionally independent, given Z, then

$E[g(X,Y)|Y = u, Z = v] = E[g(X,u)|Z = v]$ a.s. $[P_{YZ}]$.

References

References

Ash, Robert B. [1970]: BASIC PROBABILITY THEORY, John Wiley & Sons, New York.

Chung, Kai Lai [1974]: ELEMENTARY PROBABILITY THEORY WITH STOCHASTIC PROCESSES, Springer-Verlag, New York.

Çinlar, Erhan [1975]: INTRODUCTION TO STOCHASTIC PROCESSES, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Gaver, Donald P., and Gerald L. Thompson [1973]: PROGRAMMING AND PROBABILITY MODELS IN OPERATIONS RESEARCH, Brooks/Cole Publishing Co., Monterey, California.

Hillier, Frederick S., and Gerald J. Lieberman [1974]: INTRODUCTION TO OPERATIONS RESEARCH, Second edition, Holden-Day, Inc., San Francisco.

Howard, Ronald A. [1960]: DYNAMIC PROGRAMMING AND MARKOV PROCESSES, Technology Press of MIT & John Wiley & Sons, Inc., New York.

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes [1974]: INTRODUCTION TO THE THEORY OF STATISTICS, Third edition, McGraw-Hill Book Company, New York.

Karlin, Samuel, and Howard M. Taylor [1975]: A FIRST COURSE IN STOCHASTIC PROCESSES, Second edition, Academic Press, New York.

Parzen, Emanuel [1962]: STOCHASTIC PROCESSES, Holden-Day, Inc., San Francisco.

Pfeiffer, Paul E., and David A. Schum [1973]: INTRODUCTION TO APPLIED PROBABILITY, Academic Press, New York.

Rényi, A. [1970]: PROBABILITY THEORY, American Elsevier Publishing Company, Inc., New York.

Schum, David A., and Paul E. Pfeiffer [1973]: "Observer Reliability and Human Inference." IEEE Transactions on Reliability, vol R-22, no. 3, August, 1973, pp 170-176.

Schum, David A., and Paul E. Pfeiffer [1977]: "A Likelihood Ratio Approach to Classification Problems using Discrete Data", Organizational Behavior and Human Performance 19, 207-225, August, 1977.

Scott, D. W., L. Factor, and G. A. Gorry. [1978]: A Model for Predicting the Distribution of Response Time for an Urban Ambulance System (to appear).

Selected Answers, Hints, and Key Steps

Selected Answers, Hints, and Key Steps

A-1 i) $\mathcal{F}(X) = \{\emptyset, A, A^c, \Omega\}$ ii) $\mathcal{F}(X) = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, \Omega\}$

A-2 i) $X^{-1}((-\infty, 0]) = A \uplus B$ iii) $X^{-1}((-\infty, 3]) = A \uplus B \uplus C = D^c$

A-3 $g_1 = g_3$ a.s. $[P_X]$, but $g_3 \neq g_2$ a.s. $[P_X]$

A-4 a) g is cont. (draw graph), hence Borel, all t

A-8 a) $X \geq 0$. $A = \biguplus_{i=1}^{\infty} A_i$ implies $I_A = \lim_n \sum_{i=1}^{n} I_{A_i}$ implies

$\sum_{i=1}^{n} I_{A_i} X$ increases to $I_A X$. Use linearity, monotone convergence.

B-6 i) implies $P(AB|D) = P(A|D)P(B|D)$ ii) implies $P(AH|D) = P(A|D)P(H|D)$

iii) implies $P(BH|D) = P(B|D)P(H|D)$ iv) implies $P(ABH|D) = P(AB|D)P(H|D)$

B-7 576/228

B-8 b) $P(C|T_1T_2)/P(C^c|T_1T_2) = 64/99$ c) $P(C|T_1T_2^c)/P(C^c|T_1T_2^c) = 16/891$

B-9 $P(W|Q)/P(W^c|Q) = 1/3$ implies $P(W|Q) = 1/4$ $P(Q) = 1/2$

$P(W|Q^c)/P(W^c|Q^c) = 3/2$ implies $P(W|Q^c) = 3/5$

$$\frac{P(W|AB^c)}{P(W^c|AB^c)} = \frac{P(Q)P(W|Q)P(A|Q)P(B^c|Q) + P(Q^c)P(W|Q^c)P(A|Q^c)P(B^c|Q^c)}{P(Q)P(W^c|Q)P(A|Q)P(B^c|Q) + P(Q^c)P(W^c|Q^c)P(A|Q^c)P(B^c|Q^c)} = \frac{41}{74}$$

B-10 {T,B} is conditionally independent, given A, and given A^c

$P(AT) = 0.54$ $P(AT^c) = 0.12$ $P(A^cT) = P(T) - P(AT) = 0.06$ $P(A^cT^c) = 0.28$

$$\frac{P(T|B)}{P(T^c|B)} = \frac{P(AT)\,P(B|A) + P(A^cT)\,P(B|A^c)}{P(AT^c)P(B|A) + P(A^cT^c)P(B|A^c)} = \frac{342}{156}$$

B-11 b) $P(D_1) = 0.2$ $P(I|D_1) = 0.1$ $P(I^c|D_1^c) = 0.2$ $P(D_2^c|I^cD_1) = 0.96$

$P(D_2|I^cD_1^c) = 0$ implies $P(D_1^cI^cD_2) = 0$ $IC = ID_2 = \emptyset$

Hence, $P(D_1^cD_2) = P(D_1^cID_2) + P(D_1^cI^cD_2) = 0$

$$P(D_2|C) = \frac{P(D_1)P(I^c|D_1)P(D_2|I^cD_1)P(C|I^cD_1D_2)}{P(D_1)P(I^c|D_1)P(C|I^c) + P(D_1^c)P(I^c|D_1^c)P(C|I^c)} = \frac{9}{340}$$

B-12 $E_p = D_{10}D_{21}D_{31}D_{42}D_{50}D_{61}$ $\lambda_p = 3.290 > -0.201$ Classify in group 1

C-3 $E[I_A g(X)] = E[I_{AB} g(X)] + E[I_{AB^c} g(X)] = E[g(X)|AB]P(AB) + E[g(X)|AB^c]P(AB^c)$

C-7 $A = \{X^2 + Y^2 \leq 1\} = \{(X,Y) \in Q\}$ Since $Z = X$ on A, we have

$E[Z|A]P(A) = E[I_Q(X,Y)X] = 0$ (by evaluation of integral). Also

$E[Z|A^c]P(A^c) = E[I_{A^c} c] = cP(A^c)$. Hence $E[Z] = 0 + cP(A^c) = c(1 - \frac{\pi}{4})$

C-8 a) $E[X^2 + Y^2|X = t] = t^2 + E[Y^2|X = t] = (3t^2 + 4)/2 \quad 1 \leq t \leq 2$

b) $E[XY|X = t] = \frac{2}{3} t(t^2 + 2t + 4)/(t + 2) \quad 1 \leq t \leq 2$

c) $E[X|X \leq \frac{1}{2}(Y + 1)] = E[XI_Q(X,Y)]/E[I_Q(X,Y)] = \frac{93}{180} \frac{108}{47} = \frac{279}{235} \approx 1.19$

C-9 a) $E[X^2 + Y^2|X = t] = t^2 + E[Y^2] = t^2 + 7/6 \quad -1 \leq t < \infty$

b) $E[XY|X = t] = tE[Y] = t/4 \quad -1 \leq t < \infty$

C-10 $E[g(X,Y)|Y = u, Z = v] = E[g^*(X,Y,Z)|Y = u, Z = v]$

$= E[g^*(X,u,v)|Y = u, Z = v]$ by CE10)

$= E[g(X,u)|Y = u, Z = v]$

C-11 $E[g(X,Y)|Z = v] = E[e(Y,Z)|Z = v] = E[e(Y,v)|Z = v]$

$= \int e(u,v)\, dF_{Y|Z}(u|v) = \int E[g(X,Y)|Y = u, Z = v]\, dF_{Y|Z}(u|v)$

$= \int E[g(X,u)|Y = u, Z = v]\, dF_{Y|Z}(u|v)$ by Prob C-10

C-12 a) $v(Y) = E[X^2 - 2e(Y)X + e^2(Y)|Y] = E[X^2|Y] - 2e(Y)E[X|Y] + e^2(Y)$

c) $E[v(Y)] + Var[e(Y)] = E\{E[X^2|Y]\} - E[e^2(Y)] + E[e^2(Y)] - E^2[X]$

$= E[X^2] - E^2[X]$

C-13 a) $E[D|N = n]P(N = n) = E[I_{\{n\}}(N)D] = E[I_{\{n\}}(N)Y_n]$ since $D = Y_n$ on $N^{-1}(n)$. This implies $E[D|N = n] = E[Y_n] = nE[X]$

$Var[D|N = n] = E[D^2|N = n] - e^2(n) = E[Y_n^2] - E^2[Y_n]$

c) $Var[D] = E[v(N)] + Var[e(N)] = E\{NVar[X]\} + Var\{NE[X]\}$.

$Var[X]$ and $E[X]$ are constants.

C-14 a) $\varphi_D(u) = E\{E[e^{iuD}|N]\}$.

$E[e^{iuD}|N = n]P(N = n) = E[I_{\{n\}}(N)e^{iuD}] = E[I_{\{n\}}(N)e^{iuY_n}]$

$= P(N = n)\varphi_{Y_n}(u) = P(N = n)\varphi_X^n(u)$

$\varphi_D(u) = \sum_n P(N = n)\, \varphi_X^n(u) = g_N[\varphi_X(u)]$

C-15 a) $0 \le \mathrm{Var}[e(Y)]/\mathrm{Var}[X] = \mathrm{Var}[e(Y)]/\{\mathrm{Var}[e(Y)] + E[v(Y)]\} \le 1$

since $v(Y) \ge 0$ a.s.

d) Set $X^* = (X - E[X])/\sigma_X$ and $Y^* = \{g(Y) - E[g(Y)]\}/\sigma[g(Y)]$

$\rho^2[X,g(Y)] = E^2[X^*Y^*] = E^2\{E[X^*Y^*|Y]\} = E^2\{Y^*E[X^*|Y]\}$ by CE8)

$\le E[(Y^*)^2]E\{E^2[X^*|Y]\}$ by E 16)

$= E(E^2\{X - E[X]|Y\})/\mathrm{Var}[X] = \mathrm{Var}[e(Y)]/\mathrm{Var}[X] = K^2$

D-1 By CI11), $\{X,Y\}$ is conditionally independent, given Z. Hence

$E[g(X)h(Y)|Z] = E[g(X)|Z]E[h(Y)|Z] = E[g(X)]E[h(Y)|Z]$ by CE5)

D-2 i) $f_{H|W}(u|t_1, \ldots, t_n) = e^{(n-1)u} / \int_0^{t_0} e^{(n-1)u}\, du$, $t_0 = \min\{t_1, \ldots, t_n\}$

$E[H|W = t_1, \ldots, t_n] = t_0 e^{(n-1)t_0}/[e^{(n-1)t_0} - 1] - 1/(n-1)$ $n \ge 2$

ii) $E[H|W = k_1, \ldots, k_n] = (m + k)/(\lambda + n)$ $k = k_1 + k_2 + \ldots + k_n$

iii) $E[H|W = k_1, \ldots, k_n] = (n + 1)/(n + k + 2)$

D-3 $E[H] = 2/3$ $E[H|S_{10} = 8] = 8/11$ $\mathrm{Var}[H] = 2/117$ $\mathrm{Var}[H|S_{10} = 8] = 24/2783$

D-4 $E[I_M(H)D] = \sum_n E[I_{\{n\}}(N)I_M(H)Y_n] = \sum_n P(N = n)E\{I_M(H)E[Y_n|H]\}$

D-5 a) By CI11), $\{V,D\}$ is conditionally independent, given N.

$P(W \le t, N = n) = E[I_Q(D,V)I_{\{n\}}(N)] = E\{I_{\{n\}}(N)E[I_Q(D,V)|V,N]\}$

BY CI12) $E[I_Q(D,V)|V = v, N = n] = E[I_Q(D,v)|N = n] = P(D \le vt|N = n)$

$P(W \le t, N = n) = \sum_k \int I_{\{n\}}(k)P(D \le vt|N = k)\, dF_V(v)P(N = k)$

$= P(N = n) \int P(D \le vt|N = n)\, dF_V(v)$

b) $P(W \le t|N = n) = 20$ at, $0 \le t \le 1/20a$ $a^2 = n\pi/A$

c) $P(W \le t|N = n) = 1 + \frac{1}{10\ at}(e^{-15at} - e^{-25at})$ $0 \le t$, $a^2 = n\pi/A$

D-6 $p_X(26) = 0.0370$ $R(10,26) = -11.15$ $\underline{R(20,26) = -46.40}$

$R(30,26) = -35.35$ Optimum $a = 20$.

D-7 $\ell(a,u) = 1 - 2p(a,u)$, where

$p(a,u) = P(Y = a-1|H = u) + P(Y = a|H = u) + P(Y = a+1|H = u)$

$= C(n, a-1)u^{a-1}(1 - u)^{n-a+1} + C(n,a)u^a(1 - u)^{n-a}$

$+ C(n,a+1)u^{a+1}(1 - u)^{n-a-1}$

$$R(a,x) = E[\ell(a,H)P(X = x|H)]/P(X = x) \qquad P(X = x) = 1/(m+1)$$

$$= 1 - \frac{2(m+1)C(m,x)}{n + m + 1} K(a,x)$$

To minimize $R(a,x)$, maximize with respect to a the function

$$K(a,x) = \frac{C(n,a-1)}{C(n+m,a+x-1)} + \frac{C(n,a)}{C(n+m,a+x)} + \frac{C(n,a+1)}{C(n+m,a+x+1)}$$

For $n = 10$, $m = 3$, $x = 2$, $K(5,2) = 0.4324$ $K(6,2) = 0.4779$

$K(7,2) = 0.4883$ $K(8,2) = 0.4534$ $K(9,2) = 0.3625$

Optimum $R(7,2) = 1 - \frac{12}{7} K(7,2) = 0.1628$

D-8 Strategy: 1st stage-- risk $\varphi_1(0,0) = 47/12$ = expected gain for strategy

2nd stage-- If successful $\sim \varphi_2(1,1)$, then risk

If unsuccessful $\sim \varphi_2(1,0)$, then play safe

3rd stage-- $\varphi_3(2,2)$ indicates risk

$\varphi_3(2,1)$ indicates risk

$\varphi_3(1,0)$ indicates safe

E-1 b) $\{T_1 = k\} = \{U_{k-1} \in M^c \times M^c \times \ldots \times M^c \times M = M_k \subset S^{k+1}\} = A_k, \quad M = \{i\}$

c) $\{T_2 = k\} = \biguplus_{j=0} \{T_1 = j\}\{W_{j+1,k} \in M'_{k-j}\} = \{U_k \in \biguplus_{j=0} M_j \times M'_{k-j} = Q_k\}$

E-2 $E[g(Y)I_Q(U_T)] = E[g(Y) \sum_k I_{M_k}(U_k)I_{Q(k)}(U_k)]$

$= \sum_k E\{E[g(Y)|U_k]I_{M_k}(U_k)I_{Q(k)}(U_k)\}$

$= E\{\sum_k E[g(Y)|U_k]I_{M_k}(U_k)I_Q(U_T)\}$

E-3 a) From problem E-2

$$E[g(W_{T,T+n})|U_T] = \sum_k E[g(W_{k,k+n})|U_k]I_{M_k}(U_k)$$

$$= \sum_k E[g(W_{k,k+n})|X_k]I_{M_k}(U_k) \quad \text{by Markov property}$$

$$E[g(W_{T,T+n})I_M(X_T)] = \sum_k E\{E[g(W_{k,k+n})I_M(X_k)|U_k]I_{M_k}(U_k)\}$$

$$= \sum_k E\{E[g(W_{k,k+n})|X_k]I_M(X_k)I_{M_k}(U_k)\}$$

$$= E\{\sum_k E[g(W_{k,k+n})|X_k]I_{M_k}(U_k)I_M(X_T)\}$$

Hence $E[g(W_{T,T+n})|X_T] = \sum_k E[g(W_{k,k+n})|X_k]I_{M_k}(U_k)$ a.s.

E-4 a) $E[X_{n+1}] = E\{E[X_{n+1}|Y_0, Y_1, \ldots, Y_n]\} = E[X_n]$

b) $E[X_{n+1}|U_n] = E[X_{n+1} - X_n|U_n] + E[X_n|U_n]$

$= E[X_{n+1} - X_n] + X_n$ a.s. by Thm E3-2, CE5), and CE7)

$= 0 + X_n$