a mathematical theory of evidence

Glenn Shafer

LIMITED PAPERBACK EDITIONS / PRINCETON UNIVERSITY PRESS

a mathematical [redacted] theory of evidence

a mathematical theory of evidence

Glenn Shafer

Princeton University Press

Published by Princeton University Press, Princeton and London

Library of Congress Cataloging in Publication Data will
be found on the last printed page of this book

This book has been composed in Varityper Bookman
Printed in the United States of America
by Princeton University Press, Princeton, New Jersey

This book is dedicated to the memory
of Tempa Shafer and Ralph Fox,
two of my teachers.

FOREWORD

Glenn Shafer has greatly extended, refined, and recast a theory on which I expended much effort in the 1960's. The mathematical theory has been rebuilt around the concept of combining simple support functions and their corresponding weights of evidence, and many interesting new mathematical results are presented. Simultaneously, the theory has broken free of the narrow statistical confines of random sampling in which I worked, and has been reexpressed in terms of more general relevance such as belief, support, and evidence.

My own work grew out of attempts first to understand and then to replace the fiducial argument of R. A. Fisher. Beginning about fifty years ago, the British-American school believed that it had, by various pragmatic devices, banished the Bayesian scourge. But I gradually came to the perception that my revision of the fiducial argument was merely a loosening of the formalism of Bayesian inference, almost identical in purpose, yet more cautious because it assigned propositions lower probabilities (Shafer's degrees of belief) rather than simple additive probabilities.

I differ from Shafer in that I am comfortable with the view that subjective, epistemic probability is the essential concept, while chance or physical probability is only a subspecies which scientific tradition has come to regard as "objective." Moreover, I believe that Bayesian inference will always be a basic tool for practical everyday statistics, if only because questions must be answered and decisions must be taken, so that a statistician must always stand ready to upgrade his vaguer forms of belief into precisely additive probabilities. It is nevertheless very important to study theories which permit discrimination of circumstances where knowledge is secure enough to permit fair bets from other circumstances where the concept of fair bet becomes increasingly meaningless. One principle of rationality would be to confine one's enterprises to the former type of circumstance.

The mysteries of probable reasoning are unlikely ever to disappear, but the techniques, models, and formalisms in current use will undergo gradual sideways motion as well as growth. One current trend is to worry more about the prior knowledge embodied in the specification (i.e., assumed parametric model) part of the traditional models of statistics, rather than focusing entirely on the "prior" distributions of Bayesian inference. Witness the current interest in robustness studies. Shafer points out that the data represented by a specification cannot offer support for that specification, and he offers weight of conflict as a new criterion for comparing specifications. As one thought for the future, I wish to suggest that, just as traditional Bayesian reasoning has been shaken loose from its moorings, perhaps there will appear a comparable weakening of the strong form of information implied by a typical specification.

A. P. DEMPSTER

JULY 1975

PREFACE

In the spring of 1971 I attended a course on statistical inference taught by Arthur Dempster at Harvard. In the fall of that same year Geoffrey Watson suggested I give a talk expositing Dempster's work on upper and lower probabilities to the Department of Statistics at Princeton. This essay is one of the results of the ensuing effort. It offers a reinterpretation of Dempster's work, a reinterpretation that identifies his "lower probabilities" as epistemic probabilities or degrees of belief, takes the rule for combining such degrees of belief as fundamental, and abandons the idea that they arise as lower bounds over classes of Bayesian probabilities.

In writing the essay I have tried to combine mathematical rigor with an emphasis on the intuitive ideas that the mathematical definitions represent. I have provided thorough proofs for all displayed theorems, but in order not to clutter the reader's view of the main ideas, I have banished these proofs to appendices at the ends of the chapters. And in an effort to keep the mathematics manageable for both author and reader I have limited the exposition to the case where the set of possibilities one considers is finite. Thus even the proofs, though they are sometimes closely reasoned, never appeal to mathematical facts more advanced that the binomial theorem or the properties of the exponential function. This de-emphasis of the purely mathematical aspects of the essay's theory will not, I hope, deter readers from attacking the genuine mathematical challenges involved in generalizing the theory to infinite sets of possibilities and in extending it in the direction suggested in Chapter 8.

* *

I am indebted to Art Dempster personally as well as intellectually, and I would like to thank him for personally attending to my work and for consenting to write a foreword to this essay.

I would also like to thank the many friends, colleagues, and teachers who have helped me with these ideas during the past several years. Foremost among these is my wife Terry, whose comments have been helpful at every stage. The early encouragement of Peter Bloomfield, Gary Simon, and Geoffrey Watson was crucial, and at the final stage of preparing the exposition I profited particularly from comments by Frank Anscombe, Bill Homer, and Richard Jeffrey. Larry Rafsky of Bell Telephone Laboratories checked most of the proofs.

I am indebted to Princeton University and the taxpayers of the United States for financial support. I learned most of the basic ideas of this essay during the academic years 1971-73, while I was supported by a National Science Foundation graduate fellowship, and I learned more during the summer of 1973, while my salary was paid by contract N00014-67A0151-0017 from the Office of Naval Research. I began writing the essay in the spring of 1974 while teaching an undergraduate course called *Probability and Scientific Inference*, and I completed the bulk of it during the summer of 1974, with financial support from National Science Foundation grant GP-43248.

Finally, I would like to thank all those involved in the physical preparation of the book. Florence Armstrong did an excellent and timely job of typing the draft originally submitted to the Princeton University Press. And special commendation is due the dedicated individuals associated with the Press itself who make it possible to publish this and other scholarly books in this relatively inexpensive form.

GLENN SHAFER
PRINCETON, NEW JERSEY
JUNE 26, 1975

CONTENTS

FOREWORD by A. P. Dempster vii

PREFACE ix

CHAPTER 1. INTRODUCTION 3

§1. Synopsis 4
§2. The Idea of Chance 9
§3. The Doctrine of Chances 12
§4. Chances as Degrees of Belief 16
§5. The Bayesian Theory of Partial Belief 18
§6. The Role of Judgment 20
§7. The Representation of Ignorance 22
§8. Combination vs. Conditioning 25
§9. The Representation of Probable Reasoning 28
§10. Statistical Inference 29
§11. The Bayesian Theory as a Limiting Case 32
§12. Probability 33

CHAPTER 2. DEGREES OF BELIEF 35

§1. Subsets as Propositions 35
§2. Basic Probability Numbers 37
§3. Belief Functions 39
§4. Commonality Numbers 40
§5. Degrees of Doubt and Upper Probabilities 42
§6. Bayesian Belief Functions 44
§7. Mathematical Appendix 46

CHAPTER 3. DEMPSTER'S RULE OF COMBINATION 57

§1. Combining Two Belief Functions 58
§2. Multiplying Commonality Numbers 61
§3. Combining Several Belief Functions 62
§4. The Weight of Conflict 64
§5. Conditioning Belief Functions 66
§6. Other Properties of Dempster's Rule 67
§7. Mathematical Appendix 68

CHAPTER 4. SIMPLE AND SEPARABLE SUPPORT FUNCTIONS 74

§1. Simple Support Functions 75
§2. Bernoulli's Rule of Combination 75
§3. The Weight of Evidence 77
§4. Heterogeneous Evidence 79
§5. Conflicting Evidence 82
§6. Separable Support Functions 86

CHAPTER 5. THE WEIGHTS OF EVIDENCE 88

§1. Decomposing Separable Support Functions 89
§2. Combining Weights of Evidence 90
§3. The Assessment of Evidence 93
§4. The Weight of Internal Conflict 95
§5. The Impingement Function 97
§6. The Weight-of-Conflict Conjecture 98
§7. Some Numerical Examples 100
§8. Mathematical Appendix 103

CHAPTER 6. COMPATIBLE FRAMES OF DISCERNMENT 114

§1. Refinements and Coarsenings 115
§2. The Inner and Outer Reductions 117
§3. Is There an Ultimate Refinement? 119
§4. Families of Compatible Frames 121
§5. Consistent Belief Functions 125
§6. Independent Frames 127
§7. Mathematical Appendix 129

CHAPTER 7. SUPPORT FUNCTIONS 141

§1. The Class of Support Functions 142
§2. Support, Dubiety, and Plausibility 144
§3. The Vacuous Extension of a Support Function 146
§4. Evidential Independence 147
§5. Cognitive Independence 149
§6. Mathematical Appendix 152

CHAPTER 8. THE DISCERNMENT OF EVIDENCE 172

§1. Families of Compatible Support Functions 173
§2. Discerning the Interaction of Evidence 175
§3. Discerning Weights of Evidence 182
§4. If the Weight-of-Conflict Conjecture is True 185
§5. Mathematical Appendix 186

CHAPTER 9. QUASI SUPPORT FUNCTIONS 196

§1. Infinite Contradictory Evidence 197
§2. The Class of Quasi Support Functions 199
§3. Chances are not Degrees of Support 201
§4. The Bayesian Profusion of Infinite Weights 202
§5. Bayes' Theorem 204
§6. Mathematical Appendix 209

CHAPTER 10. CONSONANCE 219

§1. Consonant Support Functions 219
§2. The Contour Function 221
§3. The Embarrassment of Dissonance 223
§4. Inferential Evidence 226
§5. Mathematical Appendix 229

CHAPTER 11. STATISTICAL EVIDENCE 237

§1. A Convention for Assessing Statistical Evidence 238
§2. The Weights of Evidence 245
§3. Epistemic vs. Aleatory Combination 247
§4. The Effect of a Bayesian Prior 250
§5. Discounting Statistical Evidence 251
§6. Specifications on Compatible Frames 255
§7. The Role of Supposition 258
§8. Perspective 262
§9. Mathematical Appendix 264

CHAPTER 12. THE DUAL NATURE OF PROBABLE REASONING 274

§1. The Effect of Assumptions 274
§2. The Need for Assumptions 276
§3. Choosing our Frames of Discernment 279
§4. The Role of Epistemic Probability 284
§5. Two Tasks 285

BIBLIOGRAPHY 287

INDEX 292

a mathematical theory of evidence

A MATHEMATICAL THEORY OF EVIDENCE

Glenn Shafer

CHAPTER 1. INTRODUCTION

> By *chance* I mean the same as probability.
>
> THOMAS BAYES (1702-1761)

> Ainsi, un événement aura, par sa nature, une chance plus ou moins grande, connue ou inconnue; et sa probabilité sera relative à nos connaissances, en ce qui le concerne.
>
> SIMEON DENIS-POISSON (1781-1840)

The mathematical theory presented in this essay is at once a theory of evidence and a theory of probable reasoning. It is a theory of evidence because it deals with weights of evidence and with numerical degrees of support based on evidence. It is a theory of probable reasoning because it focuses on the fundamental operation of probable reasoning: the combination of evidence.

The theory begins with the familiar idea of using a number between zero and one to indicate the degree of support a body of evidence provides for a proposition – i.e., the degree of belief one should accord the proposition on the basis of the evidence. But unlike past attempts to develop this idea, the theory does not focus on the act of judgment by which such a number is determined. It focuses instead on something more amenable to mathematical analysis: the combination of degrees of belief or support based on one body of evidence with those based on an entirely distinct body of evidence. The heart of the theory is Dempster's rule for effecting this combination.

This introductory chapter provides a synopsis of our theory, contrasts its subject matter with the subject matter of the mathematical theory of chance, and contrasts its approach with the approach of a more familiar theory of partial belief: the Bayesian theory.

The synopsis follows in §1 below; it describes, of course, only the most salient features of the following chapters. The theory of chance is taken up in §§2-4, where we learn how chance differs from partial belief and why chances are sometimes interpreted as degrees of belief. The Bayesian theory is introduced in §5, where it is explained as an attempt to appropriate the rules for chances as rules for degrees of belief. And the Bayesian theory is contrasted with and related to our theory in §§6-11. Finally, §12 takes up the meaning of the word "probability," a semantic issue that often entangles discussions of chance and partial belief.

I discuss the Bayesian theory at length in this chapter because of the essential role this very controversial theory has played in the historical development of the idea of partial belief or "subjective probability." In the past almost all students of this idea have tied it to the Bayesian theory: those who have been committed to the value of the idea have invariably adopted and defended the Bayesian theory, while those who have rejected the Bayesian theory have tended to consider their objections to the theory proof of the inviability of the idea. As we will see in §§5-11 below, our theory frees the idea of partial belief from the Bayesian theory and develops it in a way that should appeal to both sides in the Bayesian controversy. As I explain in §5 and §§10-11, our theory includes the Bayesian theory as a special case and thus retains at least some of the attraction of that theory. And as I explain in §§6-9, the divergences between our theory and the Bayesian theory are closely related to the objections that the opponents of Bayesian theory have found so convincing.

§1. Synopsis*

Suppose Θ is a finite set, and let 2^{Θ} denote the set of all subsets

*For an explanation of the mathematical notation, see the appendix to Chapter 2.

of Θ. Suppose the function $Bel: 2^\Theta \to [0,1]$ satisfies the following conditions:

(1) $Bel(\emptyset) = 0$.

(2) $Bel(\Theta) = 1$.

(3) For every positive integer n and every collection $A_1, \cdots, A_n$ of subsets of Θ,

$$Bel(A_1 \cup \cdots \cup A_n) \geq \sum_i Bel(A_i) - \sum_{i<j} Bel(A_i \cap A_j) + - \cdots + (-1)^{n+1} Bel(A_1 \cap \cdots \cap A_n).$$

Then Bel is called a *belief function* over Θ. This essay explores the possibility of using such functions to represent partial belief.

Such a possibility arises when the set Θ is interpreted as a set of possibilities, exactly one of which corresponds to the truth. For each subset A of Θ, the number Bel(A) can then be interpreted as one's degree of belief that the truth lies in A. And rules (1)-(3) can be understood as rules governing these degrees of belief.

EXAMPLE 1.1. *The Ming Vase.* I contemplate a vase that has been represented as a product of the Ming dynasty. Is it genuine or is it counterfeit?

Let θ_1 correspond to the possibility the vase is genuine, θ_2 to the possibility it is counterfeit. Then

$$\Theta = \{\theta_1, \theta_2\}$$

is the set of possibilities, and

$$2^\Theta = \{\emptyset, \Theta, \{\theta_1\}, \{\theta_2\}\}$$

is the set of its subsets. A belief function Bel over Θ represents my belief if $Bel(\{\theta_1\})$ is my degree of belief that the vase is genuine and $Bel(\{\theta_2\})$ is my degree of belief that it is counterfeit.

Denote $s_1 = Bel(\{\theta_1\})$ and $s_2 = Bel(\{\theta_2\})$. Then rule (3) above imposes certain restrictions on the values that s_1 and s_2

can take. But these restrictions are not severe; as we will see in Chapter 2, they boil down to the requirement that $s_1 + s_2 \leq 1$. Notice that the different pairs of values (s_1, s_2) satisfying this requirement correspond intuitively to different situations with respect to the weight of evidence on the two sides of the issue. If I have little evidence on either side – little reason either to believe or disbelieve the genuineness of the vase – then I will set both s_1 and s_2 very low; in the extreme case of no evidence at all, I will set both exactly equal to zero. If, on the other hand, the evidence almost conclusively favors the genuineness of the vase, then I will set s_1 near one and s_2 near zero. Finally, substantial evidence on both sides of the issue will lead me to profess some belief on both sides; I might, for example, set $s_1 = .4$ and $s_2 = .3$. ■

There would be no sense in any claim that degrees of belief are compelled to obey rules (1)-(3). And I do not pretend that an individual would be "irrational" to profess degrees of belief that do not obey these rules. But the rules are intuitively attractive and essential to our theory. They are intuitively attractive because they derive from a simple picture – a picture in which one's belief is divisible and having a certain degree of belief amounts to committing a certain portion of one's belief. And they are essential to our theory because only those set functions that obey them can be combined by Dempster's rule of combination.

Mathematically, Dempster's rule is simply a rule for computing, from two or more belief functions over the same set Θ, a new belief function called their *orthogonal sum*. The burden of our theory is that this rule corresponds to the pooling of evidence: if the belief functions being combined are based on entirely distinct bodies of evidence and the set Θ discerns the relevant interaction between those bodies of evidence, then the orthogonal sum gives degrees of belief that are appropriate on the basis of the combined evidence.

Our exposition begins with a study of belief functions in Chapter 2 and with a presentation of Dempster's rule in Chapter 3. It then turns, in Chapter 4, to the task of illustrating Dempster's rule and developing the perspective the rule affords on the representation of evidence. The first step in developing this perspective is to identify the simple support functions.

A belief function $Bel : 2^{\Theta} \to [0,1]$ is called a *simple support function* if there exists a non-empty subset A of Θ and a number s, $0 \le s \le 1$, such that

$$Bel(B) = \begin{cases} 0 & \text{if } B \text{ does not contain } A \\ s & \text{if } B \text{ contains } A \text{ but } B \neq \Theta \\ 1 & \text{if } B = \Theta . \end{cases}$$

Such a belief function corresponds intuitively to a body of evidence whose precise and full effect is to support the subset A to the degree s. By virtue of supporting A, such evidence also supports any subset containing A. But it provides no support for the subsets of Θ that do not contain A.

When combined by Dempster's rule, two simple support functions focused on the same subset A yield another simple support function focused on A. But when we combine two or more simple support functions with different foci, we typically obtain a belief function that is not a simple support function, but is rather more complicated. Combination thus leads us from the class of simple support functions to a larger class of belief functions, a class I call the *separable support functions*.

As the nomenclature suggests, the simple and separable support functions are included in a yet larger class of belief functions called *support functions*. The support functions are easy to describe: they include all those belief functions that can be obtained by beginning with a separable support function on a certain set of possibilities and then "coarsening" that set of possibilities by neglecting to distinguish between certain of its elements. But a precise statement of what is meant by coarsening requires some care, and our study of support functions thus follows a thorough discussion, in Chapter 6, of coarsening and the opposite process of refining.

Closely related to the idea of a simple support function is a second idea also introduced in Chapter 4 – the idea of a *weight of evidence.* The degree of support s that a simple support function assigns its focus A should, intuitively, be determined by the weight w of the evidence pointing to A; s should increase as w increases, tending towards its maximum value 1 as w becomes indefinitely large. By requiring that the combination of two simple support functions on A should correspond to addition of the corresponding weights of evidence, we discover, in Chapter 4, a more precise relation:

$$s = 1 - e^{-w} .$$

Thus we may call

$$w = -\log(1-s)$$

the weight of evidence associated with the simple support function.

The description of simple support functions in terms of weights of evidence extends straightforwardly to separable support functions; as we see in Chapter 5, a separable support function over a set Θ corresponds to the specification of a weight of evidence for each proper non-empty subset of Θ. It is more problematic, and more interesting, to try to extend the idea of weights of evidence to the whole class of support functions. Indeed, since a given support function over a set Θ can always be obtained from any of many different separable support functions over various refinements of Θ, what can be said about the weights of evidence underlying it? I have not solved the mathematical problem posed by this question. But I do propose a conjecture, the "weight-of-conflict conjecture," which may offer the basis for a solution. The conjecture is stated in §6 of Chapter 5, and its implications are discussed in §§3-4 of Chapter 8.

The idea of weights of evidence also helps us understand the nature of the *quasi support functions* – the belief functions that do not qualify as support functions. As we see in Chapter 9, a quasi support function can always be obtained as the limit of a sequence of support functions. And roughly speaking, a sequence of support functions that has a quasi support function as its limit always exhibits contradictory weights of evidence tending simultaneously to infinity; the values of the quasi support function are determined by the finite values to which the differences among these contradictory weights tend. It is the intuitive dubiety of this picture of finite differences among infinite weights that motivates the name *support function* for those belief functions that can be obtained from weights of evidence without a limiting process and the name *quasi support function* for those that can be obtained only through such a limiting process.

The demonstration in Chapter 9 of the nature of quasi support functions marks the end of the essay's central thread of mathematical argument. Following this demonstration I introduce consonant support functions (Chapter 10), and discuss the problem of statistical inference (Chapter 11 and part of Chapter 12). And finally, in Chapter 12, I point out the role of assumption in the formulation of one's set of possibilities Θ and emphasize the consequent limitations on our theory.

§2. The Idea of Chance

For several centuries, the idea of numerical degree of belief has been identified, in both popular and scholarly thought, with the idea of chance. For most laymen and even many mathematicians, the two ideas are united under the name *probability*. But the reader will find the present essay intelligible only if he rejects this unification. Both numerical degrees of belief and chances have their roles to play in the following chapters, but most of the numerical degrees of belief studied there are not chances and do not obey all the rules obeyed by chances.

Chances arise only when one describes an *aleatory* (or *random*) *experiment*, like the throw of a die or the toss of a coin. The outcome of such an experiment varies randomly from one physically independent trial to another, and the proportion of the time that a particular one of the possible outcomes tends to occur is called the chance of that outcome. If $\mathfrak{X}$ denotes the set of all possible outcomes, and if one specifies the chance $q(x)$ for each possible outcome $x \in \mathfrak{X}$, then one has specified a function

$q:\mathcal{X} \to [0,1]$. If the set $\mathcal{X}$ is finite – and I shall assume throughout this essay that it is – then the function q completely specifies the chances involved in the experiment. I will call q the *chance density* governing the experiment.

In addition to satisfying $0 \leq q(x) \leq 1$ for all $x \in \mathcal{X}$, a chance density must also satisfy

$$\sum_{x \in \mathcal{X}} q(x) = 1 \; ; \qquad (1.1)$$

being proportions, the chances must add to one. But this is the only condition that a function $q:\mathcal{X} \to [0,1]$ must satisfy in order to qualify as a chance density. (In particular, q(x) may be zero for some x; such an x will not be "possible" after all.) There are, therefore, many chance densities on a given set $\mathcal{X}$, and knowledge of the set $\mathcal{X}$ of possible outcomes of an aleatory experiment hardly tells us what chance density governs that experiment.

EXAMPLE 1.2. *Dime-Store Dice.* Willard H. Longcor of Waukegan, Illinois, reported in the late 1960's that he had thrown a certain type of inexpensive plastic die over one million times, using a new die every 20,000 throws.* In order to avoid recording errors, Longcor recorded only whether the outcome of each throw was odd or even, but a group of Harvard scholars who analyzed Longcor's data and studied the effects of the drilled pips in the die guessed that the chances of the six different outcomes might be approximated by the numbers in the following table:

x	1	2	3	4	5	6	Total
q(x)	.155	.159	.164	.169	.174	.179	1.000

*See the article by Iverson, et al., *Psychometrika*, 1971.

They obtained these numbers by calculating the excess of even over odd in Longcor's data and supposing that each side of the die is favored in proportion to the extent that it has more drilled pips than the opposite side. The 6, since it is opposite the 1, is the most favored. ∎

Besides the proportion of the time that the actual outcome of an experiment tends to be a particular element x of $\mathfrak{X}$, we may also interest ourselves in the proportion of the time that the actual outcome tends to be in a particular subset U of $\mathfrak{X}$. This latter proportion is called the chance of U occurring; it may be denoted by Ch(U) and calculated by adding the chances for the various elements of U:

$$\mathrm{Ch}(U) = \sum_{x \in U} q(x) \ . \qquad (1.2)$$

The function Ch obviously conveys exactly the same information as q. I will call it the *chance function* corresponding to q.

A function $\mathrm{Ch}: 2^{\mathfrak{X}} \to [0,1]$ is a chance function – i.e., it can be obtained from some chance density q on $\mathfrak{X}$ – if and only if it obeys the following rules:

(1) $\mathrm{Ch}(\emptyset) = 0$.

(2) $\mathrm{Ch}(\mathfrak{X}) = 1$.

(3) If $U, V \subset \mathfrak{X}$ and $U \cap V = \emptyset$, then $\mathrm{Ch}(U \cup V) = \mathrm{Ch}(U) + \mathrm{Ch}(V)$.

(This is proven in Chapter 2 below – see Theorem 2.9. Notice that rule (1), which assigns zero to the actual outcome being in the empty set, accords with (1.2) by means of the mathematical convention that a sum of no terms is zero.) These three rules may be called the *basic rules for chances*. The third one, which says that the chances of disjoint sets add, is called the *rule of additivity* for chances.

As a comparison of the basic rules for chances with the rules for belief functions (§1 above) will reveal, it is the rule of additivity that I reject as a rule for degrees of belief.

§3. The Doctrine of Chances

The theory of chance – the mathematical theory that Abraham De Moivre (1667-1754) called the doctrine of chances – begins with the simple idea of a chance density $q:\mathcal{X} \to [0,1]$ describing an aleatory experiment with outcomes in $\mathcal{X}$. But its fascination and vigor derive from the idea of a *product* chance density and from the related but less important idea of a *conditional* chance density.

The idea of a product chance density arises when one contemplates a sequence of n physically independent trials of an experiment governed by a given chance density $q:\mathcal{X} \to [0,1]$. Such a sequence always results, of course, in a sequence of n elements of $\mathcal{X}$ – i.e., it results in an element $(x_1,\cdots,x_n)$ of the Cartesian product $\mathcal{X}^n$. One may therefore think of the n physically independent trials as a single aleatory experiment – a *compound experiment* that has $\mathcal{X}^n$ as its set of outcomes. The basic intuition underlying the doctrine of chances is that this compound experiment must be governed by the chance density $q^n:\mathcal{X}^n \to [0,1]$ defined by

$$q^n(x_1,\cdots,x_n) = q(x_1)\cdots q(x_n) .$$

(Since q^n satisfies

$$\begin{aligned}\sum_{(x_1,\cdots,x_n)\in\mathcal{X}^n} q^n(x_1,\cdots,x_n) &= \sum_{(x_1,\cdots,x_n)\in\mathcal{X}^n} q(x_1)\cdots q(x_n) \\ &= \left(\sum_{x_1\in\mathcal{X}} q(x_1)\right)\cdots\left(\sum_{x_n\in\mathcal{X}} q(x_n)\right) \\ &= 1 ,\end{aligned}$$

it is indeed a chance density.) This chance density is called the product chance density for n independent trials governed by q.

The chance function associated with q^n may be denoted by $Ch^n:\mathcal{X}^n \to [0,1]$; it assigns the chance

$$Ch^n(U) = \sum_{(x_1,\cdots,x_n)\in U} q^n(x_1,\cdots,x_n) = \sum_{(x_1,\cdots,x_n)\in U} q(x_1)\cdots q(x_n)$$

to a subset U of $\mathcal{X}^n$.

EXAMPLE 1.3. Since a single throw of one of Longcor's dice has the set of possible outcomes $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, the compound experiment consisting of two physically independent throws has the set of outcomes $\mathcal{X}^2$, which has 36 elements. If each single throw really is governed by the chance density q given in Example 1.2, then we would suppose a pair of physically independent throws to be governed by the corresponding product density q^2. If, for example, we were asked for the chance of a 6 followed by a 1, we would calculate the number

$$\begin{aligned} q^2(6,1) &= q(6)q(1) = (.179)(.155) \\ &= .028 . \end{aligned}$$

And for the chance of the sum of the two throws being 3 or less, we would calculate

$$\begin{aligned} Ch^2(\{(1,1),(1,2),(2,1)\}) &= q^2(1,1) + q^2(1,2) + q^2(2,1) \\ &= .073 . \ \blacksquare \end{aligned}$$

The product chance density is fruitful because it permits one to incorporate into the mathematical theory itself ideas that begin only as intuitively understood features of randomness. The idea that the chance $q(x)$ is equal to the proportion of the time that x occurs becomes, for example, the *law of large numbers* – the theorem that as the number n of trials increases the chance approaches one that the proportion of the observed outcomes $x_1,\cdots,x_n$ equal to x will be approximated to given accuracy by $q(x)$.

The idea of a conditional chance law arises when an aleatory experiment has been partly completed and its outcome thus partly determined – partly determined in the sense that it will now necessarily fall in a subset U of the original set $\mathcal{X}$ of possible outcomes. In such a situation the role of chance seems not yet entirely played out; there is still a chance event involved in the determination of which element of U the outcome will be. But how does one calculate the chance that this residual chance event will produce a particular element x of U? Intuitively, this chance is equal to the proportion of the time that such a residual chance event does result in x – i.e., to the proportion of those outcomes falling in U that finally fall equal to x. And this proportion is given by the ratio

$$\frac{q(x)}{Ch(U)} = \frac{Ch(\{x\})}{Ch(U)}$$

where q is the original chance density and Ch is its chance function.

So "conditioning" on a subset U of the set of possibilities of a chance density $q : \mathcal{X} \rightarrow [0,1]$ leads to a *conditional density*, the density $q_U : \mathcal{X} \rightarrow [0,1]$ defined by

$$q_U(x) = \begin{cases} \dfrac{q(x)}{Ch(U)} & \text{if} \quad x \in U \\ 0 & \text{if} \quad x \notin U \, , \end{cases}$$

where Ch is the chance function for q. Of course, we can "condition on U" only if $Ch(U) > 0$ – i.e., q must permit U to occur. (In this case,

$$\sum_{x \in \mathcal{X}} q_U(x) = \sum_{x \in U} \frac{q(x)}{Ch(U)} = \frac{Ch(U)}{Ch(U)} = 1 \, ,$$

so that q_U is indeed a chance density.)

The chance function associated with the conditional chance density q_U may be denoted Ch_U; it is given by

$$Ch_U(V) = \sum_{x \in V} q_U(x) = \sum_{x \in U \cap V} \frac{q(x)}{Ch(U)} = \frac{Ch(U \cap V)}{Ch(U)}$$

for all $V \subset \mathfrak{X}$. The chance $Ch_U(V)$, which may also be written $Ch(V|U)$, is called the *conditional chance* of V given U. And the formula

$$Ch(V|U) = \frac{Ch(U \cap V)}{Ch(U)}$$

is called the *rule of conditioning* for chances.

EXAMPLE 1.4. *Inverse Sampling.* A fair coin is to be tossed until two heads appear. The set $\mathfrak{X}$ of possible outcomes of this experiment is the set of all sequences of heads and tails that include exactly two heads, one of them at the end. For example: HTTH is in $\mathfrak{X}$, but HHT and THTHH are not. Since the coin is "fair," the chance density $q : \mathfrak{X} \to [0,1]$ assigns each element $x \in \mathfrak{X}$ the chance 2^{-n}, where n is the length of the sequence x. For example: $q(HTTH) = \frac{1}{16}$.

We begin the experiment: we toss the coin once and a head appears. The situation has changed; no sequence in $\mathfrak{X}$ that begins with T is now possible, and one that begins with H now has chance $2^{-(n-1)}$, where n is its length. For example, the chance of HTTH is $\frac{1}{8}$.

This transformation may be described, of course, by the process of conditioning. The chance density must be conditioned on the set U consisting of all sequences beginning with H, and since $Ch(U) = \frac{1}{2}$, this conditioning yields

$$q_U(x) = \begin{cases} 2q(x) & \text{if } x \in U \\ 0 & \text{if } x \notin U \end{cases}$$

$$= \begin{cases} 2^{-(n-1)} & \text{if } x \text{ begins with H, where n is the length of } x \\ 0 & \text{if } x \text{ does not begin with H.} \end{cases}$$ ■

Conditional chance has always played a relatively minor role in the mathematical theory of chance. Its importance has increased in recent decades, especially in the part of the theory concerned with "stochastic processes"; but its role is still small relative to the role played by the idea of product densities. As we will see shortly, "the rule of conditioning" figures most prominently not in the theory of chance but in a theory of partial belief.

§4. Chances as Degrees of Belief

The chances governing an aleatory experiment may or may not coincide with our degrees of belief about the outcome of the experiment. If we know the chances, then we will surely adopt them as our degrees of belief. But if we do not know the chances, then it will be an extraordinary coincidence for our degrees of belief to be equal to them.

The second case is the typical one. When we first conceive of an experiment as random, we typically have little idea about what chance density governs it. And though we may eventually form some opinion about the true chance density on the basis of actual observations of the experiment, we may never obtain any very exact or certain values for the true chances. Even after an immense number of observations, we may only have, as in the case of Longcor's dime-store dice, a guess about the true chances based on speculative assumptions.

Furthermore, scientific applications of the theory of chance usually turn on the fact that the chance density governing an experiment is, in the first instance, unknown. Typically, a scientist is interested in an aleatory experiment precisely because it might be governed by any one of several chance densities, each of which is associated with some hypothesis of scientific interest. In such a case observation of the experiment may provide evidence as to which chance density – and hence which hypothesis – is correct.

Chances, then, must be conceived of as features of the world. They are not necessarily features of our knowledge or belief. And it would be