Information Processing

VOLUME I

Boolean Algebra, Classical Logic,
Cellular Automata, and Probability Manipulations

First Revision October 2017

DAVID J. BLOWER

DEDICATED TO THE MEMORY
OF MY BELOVED DAUGHTER

ARIANE

Preface to the First Revision

It is a curious psychological feature of the human brain that it can look at some piece of text a hundred times and still allow some error to bypass the filters purposefully set up to catch it. I fully accept the utility of independent proof readers. Now, if only I could find some volunteers . . .

Since the initial publication of Volume I about six years ago, various annoying errors eventually made their presence known. I have taken the opportunity here in this Revised Edition to correct, to the best of my ability, what was either wrong, or needed saying in a different way.

Certain cosmetic changes were made as well; for example, I hate seeing words hyphenated at the end of lines, so I have tried to eliminate as many of these instances as I could. I also hate seeing single lines starting new paragraphs at the bottom of the page, so I have tried to eliminate these as well.

More embarrassing was my lack of attention when fact checking names on *Wikipedia*. I have now supplied the corrections for Dirichlet and Haldane.

I have double–checked the *Mathematica* code to make sure that I included any necessary corrections. The main change was to ensure consistency by changing all references to logic variables in *Mathematica* expressions from the previously employed symbols **A**, **B**, **C**, **. . .** to the logic inspired symbols **p**, **q**, **r**, **. . .**. This was done to avoid a conflict with *Mathematica* built–in symbols such as **C**, **D**, **E**, **I**, and so on.

I am operating under the tacit assumption that none of these relatively minor corrections alter the content and message of Volume I. None of the conceptual foundations discussed in Volume I have changed.

I have just recently completed a companion book to Volume I consisting purely of supplemental exercises. Should any reader be so inclined to look at this book, he or she will find amplifying material that I couldn't get to in my original Volume I. Since some of these exercises are a little more involved or challenging, I didn't feel quite as constrained by time and space in their solution as I would have by including them in the original Volume I.

Here is one example of what I mean. In the supplemental exercises, I conduct a very thorough examination of Feller's example of finding the probability for the number of accidents during the course of a week. Explicating the details to my satisfaction takes me eight pages in all. Nonetheless, the effort is worthwhile because it results in a more complete understanding of how Feller was using his counting formulas. But presenting it in Volume I would have been too overwhelming.

One of my faithful readers has insisted that I make a stab at providing an **Index**. Every author of a technical treatise soon becomes aware that this task is a lot harder than it first appears. Rather than just inserting labels in the text so that the same words are repeated in the **Index**, I have tried to put myself into the mind of a reader who vaguely recalls some interesting item he or she ran across in the book, but this wisp of a memory stubbornly resists all attempts to locate it.

My thanks go out to my brother Phil Blower and to my friend Romke Bontekoe for tracking down errors in Volume I. They provided me with detailed comments on various things they would like to see improved. I have included many of their suggestions, but due to my cantankerous disposition I also chose to ignore some things they wanted done.

David John Blower
Pensacola, Florida, USA
October 2017

Preface

These books begin the task of defining, explaining, arguing for, and, in the end, providing a rationale for *information processing*. This initial endeavor, Volume I, is concerned with the notion that an information processor is primarily engaged in making inferences. An information processor would definitely *prefer* to reach conclusions by using some form of deduction. Unfortunately, it is often thwarted in this desire by a fundamental lack of relevant information.

I view probability theory as a rigorous way of dealing with the uncertainty surrounding inference. Thus, we begin by treating some of the formal manipulation rules that crop up in probability theory. To provide some foundational basis for the applications to appear in later Volumes, the topics of Boolean Algebra, Classical Logic, and Cellular Automata will make an appearance here.

The second Volume will treat the issue of assigning legitimate numerical values to probabilities. The concept of Information Entropy will make its appearance there. Thus, several topics that can only be discussed from a rather abstract perspective here in Volume I will be dealt with more clearly in Volume II.

We shall discover there how familiar numbers get attached to probabilities. The important thing is that such numbers are seen in a new light because *information* has been actively inserted by some model proposed by an Information Processor.

Volume III will delve into Information Geometry as a mathematical justification for the concepts of Information Entropy. Subsequent Volumes will finally leverage these foundational principles towards the practical goal of applying information processing to all manner of inferential problems.

All of these developments will eventually lead to a rather different take on some of the familiar problems usually encountered in data analysis. I have learned the hard way that correct intuition about inferential problems can only be arrived at after some tough sledding through these foundational issues.

Returning to our present task, the formal manipulation rules of probability theory will be illustrated by an analogy to Boolean Algebra. Deduction can best be characterized by extending the concepts and notation of Boolean Algebra to Classical Logic. Elementary Cellular Automata will serve to alleviate the abstract

nature of Boolean Algebra as well as providing interesting examples of a purely deductive logical system. Interestingly, Cellular Automata will also be viewed from the perspective of simple examples of ontological systems.

These deductive systems are meant to be contrasted with the notion of inference. Our ultimate goal is to demonstrate that deduction is generalized by probabilistic inference. Predicting the future will be the main goal of probabilistic inference.

Why is information processing so important? There is reason to expect that the great tide of evolution will continue to sweep forward. One of the consequences is that *Homo sapiens* will eventually be replaced by more advanced entities. This may have taken place already. I will generically refer to our successors as *advanced information processors* (AIPs).

These AIPs will attempt to make optimal decisions based on their processing of information. For example, they will terraform Mars, build habitats on other planets, begin Galactic colonization, and provide virtual realities for human beings. They will progress through what has been labeled as Type I, II, and III civilizations where they will marshal, respectively, planetary, solar system, and Galactic resources.

If all proceeds according to plan, far into the future the AIPs themselves will evolve into the Omega Point. Here, the amount and type of information processing by the Omega Point will exceed our wildest comprehension. However, one fortunate consequence for us is that the Omega Point will resurrect all of us, contingent on their prodigious capacity to process information.

Independent of, and prior to any technical references, I want to acknowledge the four writers who have had the most profound impact on my intellectual life. They are: Frank J. Tipler, Stephen Wolfram, Douglas Hofstadter, and Edwin T. Jaynes. All four men, in their own idiosyncratic ways, have greatly advanced our intellectual heritage as well as making technical contributions to our understanding of information processing.

It is to Tipler that we owe our appreciation for some seminal ideas about how information processing might evolve into the far future and eventually culminate in the Omega Point. I will pay homage to these authors at varied times in my books for their truly revolutionary breakthroughs in how we think about the world.

No matter how advanced our descendants, the unseverable link joining us and them, as well as the feature that will most endear us to them, is the processing of information to build up a *state of knowledge*. And thus, we begin ...

David John Blower
Pensacola, Florida, USA
March 2011

An Author's Apologia

I am irritated by text books that do not tell me the point of their existence. Why should I invest my time and effort in trying to absorb whatever esoteric arguments the author wishes to present? Provide me, at least, with some minimal *context* for what I am about to read.

And if you are kind enough to do this much, don't do it once and hide it away somewhere where I may never read it, or forget where I read it. No reader of technical material is ever insulted by repetition. As a matter of fact, if other people possess the same frailties as I, we welcome repetition of difficult concepts. If you can explain something to me in several different ways, please take every opportunity to do so.

Tell me that reading about the details of the French Enlightenment will allow me to better understand why the Founding Fathers acted the way they did. Tell me how studying the reinforcement schedules in the Skinnerian school of psychology will allow me to better understand human behavior. Tell me why studying the electro-chemical bonds between atoms will help me understand why neurons fire the way they do. Tell me that studying quantum mechanics will reveal the ultimate nature of all reality. Tell me that trying to understand differential geometry will enable me to, in some sense, join minds with Einstein as he understood and explained gravity.

Furthermore, if it's not asking too much, please inform me as to some relevant context in which I will actually apply whatever I manage to learn. In this regard, no student has ever complained that there were too many solved numerical examples showing in detail the theoretical points made in the main part of the text.

For better or worse, human beings are made to be teleological creatures. We demand to know the purpose of things. What is the purpose of my existence? Why did they build Stonehenge? What was the point of my meeting that woman? Why do men watch football? What's that button on my remote control for?

I wish I could provide you with the answers to these questions. But you will have to be content with my fulfilling the teleological goal of explaining the purpose of these books.

What Is My Point?

I will make every effort to not commit these same sins. Here, at the outset, is my brief *apologia* for this series of books. It is my explanation to you for why you should invest your time and effort to understand my arguments. Of course, it makes sense for me to concentrate in this overview on the contents of Volume I. A similar kind of rationale will be provided in the overview to the topics occurring in succeeding Volumes.

Our goal is to understand *information processing* as it exists at this juncture in our journey to the Omega Point. In order to better understand the core ideas behind information processing, we must come to grips with some set of relevant concepts. These basic and fundamental concepts are captured in words, or phrases, such as: *degree of belief, prediction, state of knowledge, inference, epistemology, probability, deduction, ontology, frequency, uncertainty, missing information, total ignorance, models, and entropy.*

We (because none of these ideas are original with me) will eventually provide you with a complete, exhaustive, and coherent explanation unifying all of these words and concepts. And I mean that literally. There should be nothing in the final analysis that remains vague, fuzzy, or ambiguous. If any such confusion still exists about these words or their underlying conceptual meanings, then I have failed. Everything must be clear and it must be clear to you.

Let's begin by proceeding in the opposite direction from the order in which the material is presented in the book. In other words, the goals discussed first are the top–level goals. Then, we wend our way downwards to lower–level goals supporting these higher–level goals.

How does an Information Processor (IP) make a *prediction* about some future event? More precisely, how does the IP predict how many times various events will happen? An IP will thus attempt to predict the *frequencies* of things that can be observed or measured.

A prediction can only emanate from an IP that possesses a *state of knowledge* about those events it wishes to predict. Prediction depends on an IP's ability to make *inferences* based on its *state of knowledge*.

The IP's ability to make an *inference* is not the same as its ability to make a *deduction*. An *inference* is somewhat like a leap of faith because there is missing information that the IP would sorely like to know about. It seems that an IP must always uncomfortably co–exist with the uncertainty surrounding any inference that it makes.

A deduction, on the other hand, is like proving a mathematical theorem from the starting set of axioms. There is no missing information, it's all there in the set of axioms and the transformation rules, and the IP, in principle, can be certain about its conclusions.

Inference depends on *probability* where *probability* is the IP's quantified *degree of belief* that some statement is, in fact, true. *Probability* depends on *epistemology, information*, and *entropy*.

Moving even further down into the hierarchy of supporting concepts, we find that probability and inference depend on rules borrowed from formal deductive systems like Classical Logic. But it turns out that there is even a more formal, abstract, and rigorous system underlying Classical Logic. Classical Logic, it turns out, happens to be an application of Boolean Algebra.

Both Classical Logic and Boolean Algebra are examples of deductive systems. Probability generalizes Classical Logic. Thus, we must first explain what is common to Probability, Classical Logic, and Boolean Algebra. This must then be followed by whatever is new to transform a purely deductive system into an inferential system.

Another example of a deductive system is a cellular automaton. Fortunately for our cause, such cellular automata are defined in terms of Boolean Algebra and Classical Logic. They happen to be beautiful examples of dynamical, ontological, deductive systems that serve as (simplified?) stand–ins for vastly more complicated ontological models of the real world like quantum mechanics and general relativity. This is the truly remarkable argument for generalizations of cellular automata as made and defended by Stephen Wolfram.

But Wolfram claims that prediction can only be achieved by computation from a deductive system. This issue is one of the most profound (and, as yet, unanswered) questions to be tackled by this work. Does information processing add something of merit to this discussion about the ultimate nature of reality and our ability to understand and predict?

If probability generalizes deductive systems like Boolean Algebra and Classical Logic, might it not also generalize the deterministic deductive systems of cellular automata? I would like to phrase it more provocatively: Might it not also generalize any purely ontological description of the world? Thus, as I see it, information processing stands at the top of the heap. It enables an information processor to understand and predict, through its state of knowledge (epistemology), the workings of the real world (ontology).

Mathematica

In trying to explain all of this, the computer programming language *Mathematica* is an invaluable aid. We are indebted to Wolfram not only for his profound ideas about science, but on a more pragmatic level, to his development of *Mathematica*.

This language not only provides an operational method for defining concepts, but enables us to write algorithms to solve problems. It also serves as the final word on all of the many varieties of notation that will be introduced. But before we can begin to use the computer to do these sort of things, we must first understand how to do them by hand.

In the book as a whole, each Chapter tries to introduce some core concepts, present a usable notation, and solve some easy problems by hand. In order to probe into some of the more boring, but inevitably essential technical details, extensive exercises are provided at the end of each Chapter.

Several Appendices take on the further task of providing an introduction to *Mathematica*. This ancillary material supplies us with a universal notation, as well as a means for solving problems that could not be undertaken without the assistance of machine intelligence.

Solved Exercises

Another of my pet peeves is any textbook containing, as an adjunct, some enlightening problem sets. However, any solutions to these problems are completely absent. In my opinion, every single problem that the author thinks is a worthy problem for exposition should have a lengthy and detailed solution attached to it.

We are all students, for God's sake, and the reason why we are reading your book is because we don't know anything. Please help us by showing as much of your thought process as feasible as you tackle the solution to these problems you have deemed so vital to our understanding.

Invariably, it is whatever detail you might have provided us in these solutions to which we will return when we faced with similar problems that we are trying to solve on our own. Your brilliant analysis, sparkling wit, astute understanding of the subtleties of the world will all be tossed aside in our unseemly haste to find that page where you actually got down to the mechanics of solving that problem.

I pose many problems together with their detailed solutions at the end of every Chapter. This is really just another way for me to present more material that I couldn't get to in the main part of the text, or was so boring in its technical detail that it would bog you down unmercifully as you tried to slog your way to the end of the Chapter.

Nobody gets to the end of a Chapter and says to himself, "Oh goody, here are some obscure, esoteric problems the author has been thinking about for ten years and is barely able to solve himself. I think I'll spend a few minutes, after scarcely comprehending what was just explained to me in the main part of the text, and work them all out."

I have tried to achieve a balance, with varying degrees of success, of introducing in the main part of the text just enough of the mathematical point, or derivation, or symbolism to make it clear what is going on, with the ramifications relegated to the exercises. So if you feel that I have barely broached the issue in the main part of the text, you are probably right, but look carefully at the exercises to see whether I didn't include some material that gets closer to a more comprehensive treatment. My book containing a wide variety of supplemental exercises for this Volume may help as well.

More Apologies

These first several chapters on Boolean Algebra, Classical Logic, and Cellular Automata are perhaps the most abstract of anything I try to accomplish in these books, and from that perspective, in some sense the most difficult as well. I would sympathize greatly with any reader struggling through this morass of abstractness and querulously asking, "I thought you told me you were going to get to the point quickly? You are not getting to anything quickly that I can see."

Such a complaint is perfectly justified. If I could have figured out a different way of doing it, I would have. I ask the reader's indulgence in bearing with me through these introductory notions. It does start to form a coherent whole after a while, but it's like pulling the train in the World's Strongest Man Competition. It takes one heck of a lot of effort before the train starts to budge that first inch. Or, stealing a quote from the Book of *Ecclesiastes* from someone with similar concerns, "Better is the ending of the thing than the beginning thereof."

In return for your resolve to start pulling the train, I will reveal here what is at the end of the tracks. At the end of Volume I, we will understand how an Information Processor establishes its degree of belief in the number of times events will happen in the future given that it has observed the number of times these events have taken place in the past.

The practical example is this. We suppose it might be of some interest to a college admissions officer. How many students will graduate at some future point in time when we know their scores on three tests? We also know the outcomes for some number of students who have taken these tests in the past and for whom we know as well whether they successfully graduated or not.

Now, admittedly this is a rather prosaic example that would not ordinarily get your blood pumping. But such problems are placeholders for answering the, shall we say, somewhat more provocative question: How can an IP predict the future given some knowledge of the past? How can you justify any kind of inference based on various states of knowledge?

The motivation for such a question comes from Stephen Wolfram's remarkable conclusion at the end of his *A New Kind of Science*. He claims that it is impossible to predict the outcome of what will happen to our world, regardless of what state of knowledge an IP might claim to possess, except by actually simulating that world in all of its detail via an ontological model (like a cellular automaton). And it is highly unlikely that we will ever be able to compute fast enough with any such model to keep pace with how fast the Universe is actually computing our futures. Wolfram presents a compelling argument which is hard to disagree with.

Thus, any purpose evident in these books resides in the discussion of these issues. Another way of expressing my desire to explain information processing is this question: Can an argument be made that viewing the world through the lens of information processing serves as an antidote to Wolfram's pessimistic conclusions?

However, I believe that it is still an open question where hope exists for a more favorable answer. We simply have to start adding on some reasonable qualifiers.

Perhaps an information processor can voluntarily discard irrelevant information and develop a state of knowledge only about future outcomes that are relevant and important to the IP? Undoubtedly, there would be an enormous number of details that the IP couldn't predict using information processing. But what if all of these details are of no concern to the IP because, in the end, they all coalesce into some insignificant event?

Style

Many people (usually the same ones I have criticized above who eschew any attempt at providing the larger purpose or the context for what they have written) will find fault with my style of writing, or the tone that I employ. They apparently ascribe to some centralized code of conduct for technical exposition. The more bland and more obscure, the better they like it.

Another unbending rule is to never let your personality emerge through your writing. Please subjugate any strongly held opinion about anything that you might consider clearly right or clearly wrong. Never praise someone who has done a particularly masterful job of explanation, lest you be labeled as too fawning or too adulatory. On the other hand, never criticize someone who has done a horrible job of blundering, because, well it's just not the done thing. (Sometimes both praise and criticism must be directed at the same person).

Sometimes I will be blunt to the point of insensitivity. The topics addressed here seem to attract many authors who are unfortunately endowed with a shriveled sense of reasoning capacity. There exists an inordinate amount of the purest poppycock and balderdash (to use those wonderfully Victorian terms instead of the terms I might otherwise use) in many expositions of this material that are currently floating about and corrupting the minds of the innocent.

I shall attempt to adopt a positive stance and mostly comment favorably on those explanations which I deem superb. But there may be times when I shall yield to temptation and actually identify mistakes and their proponents. I indulge in this sinful pleasure more often in succeeding Volumes when we can actually pay close attention to how conventional wisdom manages to mangle certain critical conceptual issues.

Because this is such a contentious issue, let me expand a little in my own defense. The mistakes I object to are rarely computational or mathematical errors. These people are too smart to commit such basic blunders. That would have been beaten out of them in graduate school.

No; the mistakes are more often than not of the "conceptual" variety. They bungle some conceptual notion at the outset, and then proceed to build a house of (mathematical) straw, oblivious to the fact that everything they have done starting at the point of their initial confusion is a red herring.

After having made all of these horrible and scurrilous accusations against many of our so–called experts, am I not immodestly claiming for myself the self–appointed position as some kind of information processing *Übermensch*? Unfortunately, it is a sad fact that I am just as prone to making errors as the next man (women being superior creatures in any event).

As a matter of fact, I actually began to put these thoughts down on paper with the idea that most likely I was wrong and someone would set me right. But I am a stubborn person. Any argument that is going to change my mind must be a rational argument (what a bother!). So, by all means, please inform the world of any and all my errors in this exposition. I sincerely believe there must be many.

But as time went by, there were no such public pillories. Any complaints were always of a non–rational nature based on blind adherence to some school or another, or sometimes simply because, "I am recognized as an expert in this field and I say so," but never actually identifying any errors accompanied by coherent explanations as to why they were errors.

Nevertheless, from my standpoint, in either case, I win. If I am right, then I am very pleased to have been so fortunate to have stumbled onto the right way of doing things. If errors are pointed out to me, then I am also happy because I have learned something new.

But in all the years I have placed these arguments before the general public, I am still waiting. No one has yet been able to pinpoint the faults that bring everything down in ruin. Of course, everyone knows that what Fate has in store for you is not that you are recognized for a fraud and exposed to great shame and humiliation. No, far more painful is that you are simply ignored.

Some of the more charitable descriptions of my writing style might be a list like "discursive," "conversational," "chatty," "digressive," "long-winded," "aggressive," "confrontational," or, God forbid, even "polemical." Too bad. I am who I am and this is the way I write.

I guess my attitude is this. If you don't happen to like the way I write, that is your perogative. But I am not changing to suit your whims. Then, it is up to you to write your own books the way you prefer, and let the chips fall where they may.

Contents

List of Figures

List of Tables

Chapter 1

Boolean Algebra

1.1 Introduction

We begin with Boolean Algebra partly because it is a beautiful example of a formal mathematical system. Here, mathematical abstractness and generality are not the curse that we have all come to expect, but rather a blessing. The notation, allowable operations, and even the proof of theorems are all really quite easy to grasp when we restrict ourselves to the confines of a Boolean Algebra.

Moreover, the fundamental principles presented in this first Chapter carry over quite directly to Classical Logic and Elementary Cellular Automata. Our ultimate objective is, of course, to armor an Information Processor with a defensible rationale when it must reason in uncertain situations. The rigorous treatment of the resulting inferences rests on probability theory. Our intent is to show that the formal rules of probability theory share much in common with Boolean Algebra.

A great deal of emphasis is placed on the conceptual distinction between the formal aspects of manipulating abstract probability symbols and the assignment of legitimate numerical values to these probabilities. *Formal* here simply refers to operations that can be carried out irrespective of whatever numerical values might have been assigned to a probability via some model. As we shall discuss in exhausting detail later in Volume II, any numerical values assigned to abstract probabilities are conditioned on the information resident in some model.

These formal operations on abstract probabilities inherit most of their properties from Classical Logic and ultimately from Boolean Algebra. Note that a Boolean Algebra is a self–contained *deductive* system where there is no need to employ any actual numbers. There are formulas and functions in Boolean Algebra, but no numerical results. This is exactly what we want as a foundation for probability symbol manipulation because such a foundation should not be concerned with actual numerical computations.

Having introduced the notion of a Boolean Algebra, it is then straightforward to delve directly into a discussion of logic functions and Classical Logic. And one particularly interesting application combining both Classical Logic and Boolean Algebra is the notion of a Cellular Automaton.

I try to introduce the absolutely minimal amount of supporting infrastructure before presenting the major conceptual points of a Chapter. In the beginning, this necessitates some slow going because we haven't built up anything yet. One has to accept on faith that these preliminaries, inevitably quite boring in and of themselves and seemingly detached from any ultimate objective, are indeed leading up to something significant.

In the next sections, therefore, we have to start laying down the infrastructure by defining successively a Boolean Algebra, then functions in general, followed by Boolean functions in particular. Finally, we show how these Boolean functions can be expanded into so-called canonical forms. We will make extensive use of a particular canonical form called the *disjunctive normal form* in later Chapters on logic functions and cellular automata.

We also take a brief excursion into how proofs are derived within a Boolean Algebra. One begins to wonder where all of this abstractness is leading. The resulting lemmas and theorems do make an appearance when we show that Classical Logic is a deductive system just like Boolean Algebra, and the theorems involving logic functions are exactly like the ones proven within Boolean Algebra.

As mentioned, there is a certain pristine beauty to proving theorems in Boolean Algebra. Because this effort is detached from numbers, geometrical constructions, and functional symbols that we are usually exposed to when we are taught proofs in our mathematical education, we can begin to gain some appreciation for really how simple abstract mathematics can be at its core.

In addition to the stark simplicity of theorem construction, Boolean Algebra affords a wonderful introduction to axioms. All of our lemmas and theorems must, of course, begin with some set of fundamental axioms. An axiomatic structure is the characterizing feature of modern mathematics, and it underlies what we have almost unconsciously absorbed as defining algebraic operations.

It is fun to see how the basic axioms of an algebraic system like **Commutativity, Associativity** and **Distributivity** that we learned about in school and promptly ignored as boring irrelevancies to our lives do play a fundamental role in information processing.

Now on to Boolean Algebra, together with a slow, careful analysis of the role it plays in leading up to probability, inference, and information processing.

1.2   Example of a Boolean Algebra

Before we begin to list some of the axioms taken over from Boolean Algebra that are directly useful in the formal manipulation of probability symbols, we study a simple example of a Boolean Algebra. The definition of a Boolean Algebra begins by considering an abstract organization of five elements called a quintuple,

$$(\mathbf{B}, \circ, \bullet, F, T)$$

consisting of a set $\mathbf{B}$, called the carrier set, and the symbols $\circ$ and $\bullet$ referring to the two binary operations on the elements of $\mathbf{B}$. The elements F and T are special and distinct members of $\mathbf{B}$.

We wish to avoid some of the more common notational choices made in various presentations of Boolean Algebra. We do this to help eliminate the almost certain confusion between the abstract nature of a Boolean Algebra and the more familiar numerical computations using these same symbols.

For example, the notation $\circ$ is employed for the first binary operator instead of, say, the symbol $\times$. Similarly, the notation $\bullet$ is used for the second binary operator instead of the symbol $+$. Furthermore, the notation of F and T is chosen instead of the more typically seen 0 and 1. We want to highlight the extreme abstractness of the Boolean Algebra. Thus, we would rather write $x \circ T = x$ rather than $x \times 1 = x$.

In the following example, the carrier set $\mathbf{B}$ consists of the four abstract elements,

$$\mathbf{B} = \{a, a', F, T\}$$

where, in addition to the special elements of F and T, we have selected an element a and its complement a' to be part of $\mathbf{B}$. F is the complement of T and T is the complement of F.

As already mentioned, some expositions of Boolean Algebra will use the symbols 0 and 1 instead of F and T. I have avoided that notation, as helpful as it is, to hopefully eliminate the confusion about Boolean Algebra and ordinary numerical computations. Real numbers return when we consider probability functions.

The binary operations on the elements of $\mathbf{B}$ are defined by the following two tables appearing at the top of the next page. The first of these two tables, Table 1.1, shows the binary operation represented by " $\circ$ ", while the second table, Table 1.2 shows the binary operation represented by " $\bullet$ ".

These binary operators are meant to be abstract mathematical operators and we should not attach the familiar meaning of "multiplication" and "addition" to these symbols, although they certainly are analogous. Later on, the properties reflected in these two tables will be recast as a set of axioms. But for right now, we show a few examples from the tables to give the flavor of what is upcoming.

Table 1.1: *The definition of the binary operator " ∘ " in a Boolean Algebra by an operator table.*

∘	a	a'	F	T
a	a	F	F	a
a'	F	a'	F	a'
F	F	F	F	F
T	a	a'	F	T

Table 1.2: *The definition of the binary operator " • " in a Boolean Algebra by an operator table.*

•	a	a'	F	T
a	a	T	a	T
a'	T	a'	a'	T
F	a	a'	F	T
T	T	T	T	T

The syntactical form of the following expressions using ∘ and • is given in the so-called "infix" form of an operator where the operator symbol appears between the two arguments. Of course, this is what we are most familiar with since we write $2 + 3 = 5$ and $2 \times 3 = 6$ with infix operators like $+$ and $\times$.

From Table 1.1, some properties of the ∘ operator are,

$$a \circ a' \;=\; F$$

$$a' \circ F \;=\; F$$

$$T \circ a' \;=\; a'$$

and from Table 1.2, some properties of the • operator are,

$$a \bullet F \;=\; a$$

$$F \bullet F \;=\; F$$

$$T \bullet a \;=\; T$$

1.3   Definition of a Function

These binary operators $\circ$ and $\bullet$ are actually functions. Thus, only two functions are needed to define a Boolean Algebra. The following definitions lay down the preliminaries that eventually will be used to set up Boolean functions.

Definition 1 *A function f from a set S into a set T is written*

$$f : \; S \to T$$

where the function assigns to every element $x \in S$ an element $f(x) \in T$.

Definition 2 *A binary operation like $\circ$ or $\bullet$ on a set S is a function from $S \times S$ into the set S.*

$$\circ : S \times S \to S$$

Definition 3 *$S \times S$ is the* **Cartesian Product** *(direct, or cross-product) of the set S with itself.*

$$S \times S = \{ \, (x, y) \mid x \in S \text{ and } y \in S \, \}$$

that is, (x, y) is an ordered pair where both x and y come from S.

Since we are using the notation of **B** for the carrier set of a Boolean Algebra, we would write a general template for a binary operator as,

Definition 4 *The binary operator $\circ$ is defined as,*

$$\circ : \mathbf{B} \times \mathbf{B} \to \mathbf{B}$$

or, alternatively, now choosing $\bullet$ as the binary operator,

Definition 5 *The binary operator $\bullet$ is defined as,*

$$\bullet : \mathbf{B}^2 \to \mathbf{B}$$

In the first example presented above, there are ordered pairs such as (a, a) and (T, a') that belong to $\mathbf{B} \times \mathbf{B}$ or, alternatively, $\mathbf{B}^2$. A function assigns an element of $\mathbf{B}$ to all possible ordered pairs in $\mathbf{B}^2$. Tables 1.1 and 1.2 document all the assignments of the two functions $\circ$ and $\bullet$.

For example, the first cell of the $\circ$ operator table shows the functional assignment $a \circ a = a$ where the ordered pair (a, a) from $\mathbf{B}^2$ is assigned a from $\mathbf{B}$ by the binary operator $\circ$. The last cell of the $\bullet$ operator table shows the functional assignment $T \bullet T = T$ where the ordered pair (T, T) from $\mathbf{B}^2$ is assigned T from $\mathbf{B}$ by the binary operator $\bullet$.

As mentioned, by placing an operator like ∘ or • between its two arguments, we are using an "infix" notation, instead of the usual functional notation like $f(T, T)$. *Mathematica* expressions take advantage of several different notations depending on the situation and might use a "prefix" notation such as `Plus[2, 1]`, or an "infix" notation like `2 + 1`.

In the next Chapter, we will investigate the case of Classical Logic from the perspective of Boolean Algebra. Here there can be 16 different functions for,

$$f : \mathbf{B} \times \mathbf{B} \to \mathbf{B}$$

but, in the end, only two functions, analogous to ∘ and •, are all that is needed to represent the 16 functions.

1.4   An Example of a Boolean Function

We will use the Boolean Algebra presented in the opening section. Let the carrier set **B** consist of four elements,

$$\mathbf{B} = \{a, a', F, T\}$$

the same set as described before. The special elements F and T are required in a Boolean Algebra.

Even this simple example points out that, in general, Boolean Algebra does not have to concern itself solely with *indicator variables*, that is, variables that only take on the values F or T. The coming example, as simple as it is, is more general than our future applications to probability theory where all the variables *are* indicator variables.[1]

To continue, the domain $\mathbf{B} \times \mathbf{B}$ consists explicitly of the 16 elements,

$$\mathbf{B} \times \mathbf{B} \equiv \{(a, a), (a, a'), (a, F), \cdots, (T, T)\}$$

These 16 elements in $\mathbf{B} \times \mathbf{B}$ are all the possible ordered pairs of the four elements in **B**.

Each one of these elements in $\mathbf{B} \times \mathbf{B}$ can be associated with any one of the elements of **B**, as we know from the definition $f : \mathbf{B}^2 \to \mathbf{B}$. For example, (a, F) as a legitimate ordered pair from $\mathbf{B}^2$ might be assigned the value of a' from **B**. This is written as $f(a, F) = a'$.

There are very many possible functions in this Boolean Algebra. In fact, there are $4^{4^2} = 4,294,967,296$ possible functions. For the purposes of this example, we are going to look at one particular function as defined by the function table as shown at the top of the next page as Table 1.3.

---

[1] Some texts refer to this specific case as *Switching Theory* where obviously the mind–set is oriented to the design of logic circuits.

Table 1.3: *Boolean functional assignment table for two variables.*

Row	x	y	$f(x, y)$
1	a	a	F
2	a	a'	a
3	a	F	a
4	a	T	F
5	a'	a	a'
6	a'	a'	T
7	a'	F	T
8	a'	T	a'
9	F	a	F
10	F	a'	a
11	F	F	a
12	F	T	F
13	T	a	a'
14	T	a'	T
15	T	F	T
16	T	T	a'

Notice that the two variables are generically labeled as x and y. Each variable can be assigned only one of the four elements in the set **B**. The last column lists the functional assignment which, naturally enough, is labeled as $f(x, y)$. This Boolean function assignment, as detailed in the above table, is calculated according to a Boolean *formula*,

$$f(x, y) = (a' \circ x) \bullet (a \circ y')$$

An arbitrary function like $f(x, y)$ can be constructed from two other functions, our already familiar $\circ$ and $\bullet$ functions.

Take the first row where $x = a$ and $y = a$, substitute these specific values for the generic variables into the Boolean formula defining the function, and then refer back to Tables 1.1 and 1.2 to determine the result of the binary operation.

We might as well take the plunge and consider this to be the beginning of working through some formal manipulation rules. Let me repeat the warning issued before. This is not fun! It requires you to drastically slow down your reading pace, and follow along with pencil and paper.

$$
\begin{aligned}
f(a, a) &= (a' \circ a) \bullet (a \circ a') \\
&= F \bullet F \\
&= F
\end{aligned}
$$

Since that was so enjoyable, take a look at row 7 in the above table where $x = a'$ and $y = F$. For these arguments, the formula works out to,

$$
\begin{aligned}
f(a', F) &= (a' \circ a') \bullet (a \circ F') \\[2mm]
&= a' \bullet (a \circ T) \\[2mm]
&= a' \bullet a \\[2mm]
&= T
\end{aligned}
$$

The complement of F, written as F', is T, and the results of the binary operations are found as before by reading off the correct cell of the appropriate operator table.

As a final check on our understanding of the mechanical steps involved, take the last row where $x = T$ and $y = T$,

$$
\begin{aligned}
f(T, T) &= (a' \circ T) \bullet (a \circ T') \\[2mm]
&= a' \bullet (a \circ F) \\[2mm]
&= a' \bullet F \\[2mm]
&= a'
\end{aligned}
$$

This two variable example illustrates quite nicely the potential generality of a Boolean Algebra. Also, it illustrates the type of computing that is done without any numbers entering the picture. There is a certain appeal to this notion of performing operations of a completely symbolic nature without the interference of numbers.

1.5   Canonical Forms

We just examined a Boolean function that was represented by a Boolean formula. A Boolean function may be represented by an infinite number of Boolean formulas. As a consequence, there exist other equivalent Boolean formulas for the function defined by $f(x, y) = (a' \circ x) \bullet (a \circ y')$ in the last section's example. There is some merit then in seeking out from the infinite number of formulas, some formulas that have attractive characteristics.

These special formulas are called *canonical* formulas. Such canonical formulas will be used in the next Chapter when we want to represent many different functions in terms of just the analogs to the $\circ$ and $\bullet$ operators.

As a prime example of a canonical formula, we shall present the *disjunctive normal form* (DNF), or, as it is sometimes called, the *minterm canonical form*. We will be seeing quite a lot of formulas expressed in the disjunctive normal form as we progress through this Volume.

Continue to refer back to the functional table, Table 1.3, in the last section. For any two variable Boolean function, the DNF is derived from a recursive application of Boole's Expansion Theorem, stated formally as,

Boole's Expansion Theorem: *If* $f : \mathbf{B}^n \to \mathbf{B}$ *is a Boolean function, then,*

$$f(x_1, x_2, \cdots, x_n) = [\, x_1 \circ f(T, x_2, \cdots, x_n) \,] \bullet [\, x_1' \circ f(F, x_2, \cdots, x_n) \,]$$

for all $(x_1, x_2, \cdots, x_n)$ *in* $\mathbf{B}^n$.

There is a slight change in notation in the statement of this theorem. To cover the general case where there might be n variables in all, $x_1, x_2, \cdots, x_n$ is used instead of, say, (x, y) for $n = 2$, or (x, y, z) for $n = 3$. The important consequences of this theorem will now be illustrated through some examples, as well as in many places yet to come.

For a Boolean function with $n = 2$ variables, the expansion dictated by the theorem looks like this,

$$f(x, y) =$$

$$[\, f(T, T) \circ x \circ y \,] \bullet [\, f(T, F) \circ x \circ y' \,] \bullet [\, f(F, T) \circ x' \circ y \,] \bullet [\, f(F, F) \circ x' \circ y' \,]$$

See Exercise 1.10.25 for further details.

The expansion of the original formula via the DNF is seen to consist of four terms. There is a discernible pattern in these four terms of the DNF because the variables x and y appear in combinations of uncomplemented and complemented forms.

Refer back to Table 1.3 to find the values for all four expressions $f(\star, \star)$,

$$f(T, T) = a'$$

$$f(T, F) = T$$

$$f(F, T) = F$$

$$f(F, F) = a$$

Substituting, this yields the DNF for the original formula,

$$f(x, y) = (a' \circ x \circ y) \bullet (T \circ x \circ y') \bullet (F \circ x' \circ y) \bullet (a \circ x' \circ y')$$

The DNF on the right hand side can be simplified further because of the axioms that a Boolean Algebra obeys. For example, the third term, $F \circ x' \circ y$, must eventually reduce to F, and then be eliminated because of the properties involving $\circ$ and F, while the second term, $T \circ x \circ y'$ must eventually reduce to $x \circ y'$.

Thus, we could write the DNF in a shortened version as,

$$f(x, y) = (a' \circ x \circ y) \bullet (x \circ y') \bullet (a \circ x' \circ y')$$

Eventually, of course, even this shortened version would have to be proved equal to the original formula,

$$(a \circ x' \circ y') \bullet (x \circ y') \bullet (a' \circ x \circ y) = (a' \circ x) \bullet (a \circ y')$$

1.6   The Axioms for Boolean Algebra

This section contains a more systematic listing of the properties that a Boolean Algebra satisfies. These codify in more generality what we discovered by setting up the binary operator tables in a previous section. These properties have been given traditional labels so that we can refer to them more easily.

We use the notation x, y, z for three variables. We begin with the easier axioms.

Axiom 1 (Idempotence)

$$x \circ x \ = \ x \tag{1.1}$$

$$x \bullet x \ = \ x \tag{1.2}$$

Axiom 2 (Special Elements)

$$x \circ T \ = \ x \tag{1.3}$$

$$x \bullet T \ = \ T \tag{1.4}$$

$$x \circ F \ = \ F \tag{1.5}$$

$$x \bullet F \ = \ x \tag{1.6}$$

Axiom 3 (Commutativity)

$$x \circ y \ = \ y \circ x \tag{1.7}$$

$$x \bullet y \ = \ y \bullet x \tag{1.8}$$

Thus, **Idempotence** reflects the diagonal entries of Tables 1.1 and 1.2 such as $a' \circ a' = a'$, or $a \bullet a = a$. **Special Elements** permits us to fill in the entries involving F or T, such as $a \circ T = a$, or $a' \bullet F = a'$. **Commutativity** permits the "mirror–image" cell of the table to be filled in. Since $a \bullet T = T$ by **Special Elements**, $T \bullet a = T$ by **Commutativity**.

The next axiom defines the relationship between an element and its complement. For any variable x in **B**, there corresponds an element x' in **B** such that,

Axiom 4 (Complementation)

$$F' \;=\; T \tag{1.9}$$

$$T' \;=\; F \tag{1.10}$$

$$x \circ x' \;=\; F \tag{1.11}$$

$$x \bullet x' \;=\; T \tag{1.12}$$

The following association and distribution axioms are common to all algebraic systems. They tell us how elements can be shifted within parentheses and how elements are distributed across operators. For any three variables in **B**, we have,

Axiom 5 (Associativity)

$$x \circ (y \circ z) \;=\; (x \circ y) \circ z \tag{1.13}$$

$$x \bullet (y \bullet z) \;=\; (x \bullet y) \bullet z \tag{1.14}$$

Axiom 6 (Distributivity)

$$x \circ (y \bullet z) \;=\; (x \circ y) \bullet (x \circ z) \tag{1.15}$$

$$x \bullet (y \circ z) \;=\; (x \bullet y) \circ (x \bullet z) \tag{1.16}$$

We will illustrate the **Associativity** and **Distributivity** properties in the next section when we discuss three variable functions. There are other less self–evident properties of a Boolean Algebra that are sometimes listed along with these axioms just presented. But we will defer these other axioms to a later Chapter when we discuss how these axioms are translated into probability symbol manipulation axioms.

These axioms of Boolean Algebra are carried over intact to the special case of the generic probability function attached to a proposition. So we say that probability symbol manipulation inherits all of these axioms from the more general parent that is Boolean Algebra.

Probability symbol manipulation adds two important properties of its own, the **Product Rule** and the **Sum Rule**. Together, these define all that can be done in proving theorems from the strictly formal perspective of symbol manipulation.

1.7 n-variable Boolean Functions

There is no conceptual difficulty in moving up from these two variable examples to n-variable functions. The notation for an n-variable Boolean function follows directly from our experience with the $n = 2$ example.

Definition 6 *A n-variable Boolean function is defined as* $f : \mathbf{B}^n \to \mathbf{B}$

We could contemplate constructing a Boolean function of three variables from the same set $\mathbf{B} = \{a, a', F, T\}$ that we have been using all along. From the definition,

$$f : \mathbf{B}^3 \to \mathbf{B}$$

there would be some functional assignment $f(x, y, z)$ for the three arguments. An ordered triple such as (a, T, a') arises from $\mathbf{B} \times \mathbf{B} \times \mathbf{B}$. One possibility for the functional assignment might be $f(a, T, a') = F$.

Just as in the two variable case, where the variables took on the generic labels of x and y, the variables for the $n = 3$ case will be labeled x, y, and z. It is important to remember that even though the number of variables has increased, all operations remain binary operations. That is, the operations defined by the $\circ$ and $\bullet$ tables work on only two variables at a time. There now arises a need for parentheses to keep the order of operations clear. Thus, $x \circ y \bullet z$ might be $(x \circ y) \bullet z$ or $x \circ (y \bullet z)$.

1.7.1   Formal operations on three variable formulas

Two of the axioms previously listed illustrate the permissible operations using parentheses. We can present examples of the **Associativity** and **Distributivity** axioms since we now have three variables, x, y, and z.

Here is an example of the **Associativity** axiom,

$$(x \circ y) \circ z = x \circ (y \circ z)$$

Some row of the function table for this new three variable case would eventually get around to listing this possible variable assignment, $x = a, y = T, z = a'$. Does

$$(a \circ T) \circ a' = a \circ (T \circ a')?$$

We can see with the parentheses that even though we now have three variables, the operators operate on only two arguments at a time. The left and right hand sides of this equation are indeed both equal to F as shown next.

$$a \circ T \;=\; a$$

$$a \circ a' \;=\; F$$

$$T \circ a' \;=\; a'$$

$$a \circ a' \;=\; F$$

Here is an example of the **Distributivity** axiom,

$$x \circ (y \bullet z) = (x \circ y) \bullet (x \circ z)$$

For a different possible variable assignment of, say, $x = T$, $y = a'$, and $z = a$, the left and right hand sides are both equal to T.

$$T \circ (a' \bullet a) = (T \circ a') \bullet (T \circ a)$$

$$a' \bullet a = T$$

$$T \circ T = T$$

$$T \circ a' = a'$$

$$T \circ a = a$$

$$a' \bullet a = T$$

1.7.2 DNF expansion of a three variable function

Suppose that some function with three arguments is defined by a Boolean formula found through the DNF expansion. And suppose further that the expansion is a particularly simple one as in $f(x, y, z) = x \circ y \circ z$. This means that $f(T, T, T) = T$ and all the other functions involving F and T as arguments, $f(\star, \star, \star)$, must equal F. Thus, the DNF expansion is written out in full as,

$f(x, y, z) =$

$$[f(T, T, T) \circ x \circ y \circ z] \bullet [f(T, T, F) \circ x \circ y \circ z'] \bullet \cdots \bullet [f(F, F, F) \circ x' \circ y' \circ z']$$

The first term is the only term that survives,

$$f(x, y, z) = T \circ x \circ y \circ z = x \circ y \circ z$$

Above, we said that $f(a, T, a') = F$ for some function. Suppose that such a function is expressed by the Boolean formula $f(x, y, z) = x \circ y \circ z$. We could find the functional assignment for any one of the 64 rows of the function tables analogous to Table 1.3 by referring to this formula. Thus, the row for $x = a, y = T, z = a'$ would have for the final column,

$$f(x, y, z) = a \circ T \circ a' = (a \circ T) \circ a' = a \circ a' = F$$

Here is another example of a different formula for a three variable Boolean function generated via the DNF approach. In order to generate a function by the DNF process, construct a *truth table* consisting of 2^n entries of T and F. For the

current discussion centering on three variable Boolean functions, the truth table in Table 1.4 will have $2^3 = 8$ rows.

We repeat the heuristic explanation for the DNF expansion. Wherever a T appears in the second column, the corresponding variable is shown uncomplemented. Wherever an F appears, the corresponding variable is shown complemented. This construction is called a term. If the final column showing $f(x, y, z)$ has a T, then that term is included in the DNF formula. If the final column has an F, then that term is not included.

Table 1.4: *Truth table to determine a three variable function.*

Row	xyz	$f(x, y, z)$
1	TTT	T
2	TTF	F
3	TFT	F
4	TFF	F
5	FTT	T
6	FTF	F
7	FFT	F
8	FFF	F

Thus, for the above truth table, the function produced by the DNF process is,

$$
\begin{aligned}
f(x, y, z) &= (T \circ x \circ y \circ z) \bullet (T \circ x' \circ y \circ z) \\
&= (x \circ y \circ z) \bullet (x' \circ y \circ z)
\end{aligned}
$$

This formula can be reduced even further by using the axioms.

$$
\begin{aligned}
(x \circ y \circ z) \bullet (x' \circ y \circ z) &= y \circ ((x \circ z) \bullet (x' \circ z)) \\
&= y \circ (z \circ (x' \bullet x)) \\
&= y \circ (z \circ T) \\
&= y \circ z
\end{aligned}
$$

There is no reason why the truth table for another function couldn't have an a or a' in the final column in addition to the Ts and Fs. Table 1.5 is another truth table illustrating this.

Thus, the DNF expansion for this function looks like,

$$
f(x, y, z) = (axyz) \bullet (a'xy'z) \bullet (x'y'z')
$$

Table 1.5: *Another truth table to determine a different three variable function.*

Row	xyz	$f(x,y,z)$
1	TTT	a
2	TTF	F
3	TFT	a'
4	TFF	F
5	FTT	F
6	FTF	F
7	FFT	F
8	FFF	T

There are three terms corresponding to the non–F rows and the function notation has been simplified by assuming that an implicit ∘ operator exists when the variables are written together.

In Chapter Two, similar n-variable function tables will be created for the special case of the Boolean Algebra where **B** consists solely of indicator variables. This is, of course, the situation for both Classical Logic and circuit analysis where the variables can only take on values of TRUE or FALSE and the functional assignment to these variables must also be TRUE or FALSE.

And later in Chapter Three, we will explain elementary cellular automata as just further examples of an n-variable Boolean function. In particular, we will look closely at a cellular automaton made famous by Stephen Wolfram. This is his so-called Rule 110 which he showed to be a Universal Turing Machine. The rule which this cellular automaton uses to evolve over time is simply a three variable Boolean function over the carrier set **B** = {TRUE, FALSE}, and captured by the following DNF expansion of that three variable Boolean function,

$$f(x,y,z) = (xyz') \bullet (xy'z) \bullet (xy'z') \bullet (x'yz') \bullet (x'y'z)$$

1.8   Formal Manipulation Rules

We show a few symbol manipulation exercises for Boolean Algebra that will be useful for analogous probability symbol manipulations in later Chapters. We begin with some very easy transformations that apply just one or a couple of the axioms of Boolean Algebra as introduced here in Chapter One.

From this humble beginning, we build up to *lemmas*. Then, the various lemmas that have been proved, together with the beginning axioms, serve as the basic building blocks for constructing *theorems*. Finally, previously proved theorems, lemmas, and the axioms all go into the mix to prove perhaps more complicated and unanticipated theorems as consequences of the beginning set of axioms.

Just a few lemmas are presented here in the main part of the text to give you the flavor of these kind of formal manipulation rules. They are not the kind of thing you are likely to look forward to with gleeful anticipation, and attention quickly wanes after working through a couple. Several more of these kinds of proofs involving the axioms of Boolean Algebra to build up further lemmas and theorems are relegated to exercises where you can peruse them at your leisure.

The approach used here attempts to follow Wolfram's illuminating suggestions that demystify theorem proving. In particular, we derive theorems in a Boolean Algebra by applying allowable transformations, in either the forward or backward direction, to some beginning expression. This results in some further expression where other transformations are applied in a purposeful effort to reach some desired ending expression. It's all merely symbol substitution without the need for any more esoteric rationale like so called "rules of inference."

We show the parentheses for the purpose of indicating the explicit ordering of the two abstract binary operations $\circ$ and $\bullet$. The transformation applied to the previous expression is shown to the right of the current expression. In other words, as shown in Lemma 1 given below, the **Idempotence axiom**, Equation (1.1), is applied to the top expression. The final line shown, after all the transformations have been applied, is the lemma or theorem.

This first set of lemmas and theorems as derived here and in the exercises are helpful for future use in the probability symbol manipulation of joint statements where statements are repeated, and/or might occur in any order. Thus, these theorems can be used to simplify expressions and place them in some sort of natural order. It is easier to discern their import by writing out patterns like `AAA` $\rightarrow$ `A` or `BBBAA` $\rightarrow$ `AB`. These patterns indicate that, for example, the final result `A` can be obtained from initial data `AAA`.

Lemma 1 $(x \circ x) \circ x \rightarrow x$

$$(x \circ x) \circ x \qquad \text{Given}$$

$$x \circ x \qquad \text{Idempotence}$$

$$x \qquad \text{Idempotence}$$

`AAA` $\rightarrow$ `A`

Lemma 2 $((x \circ x) \circ x) \circ x \to x$

$$
\begin{array}{rl}
((x \circ x) \circ x) \circ x & \text{Given} \\
x \circ x & \text{Lemma 1} \\
x & \text{Idempotence}
\end{array}
$$

`AAAA` $\to$ `A`

Lemma 3 $(\cdots ((x \circ x) \circ x) \circ x \to x$

By induction on the first two lemmas, there exists another lemma, for applying n applications of the binary operator $\circ$ on a variable and returning that variable. The notation here is,

$$
\overbrace{\texttt{AAA} \cdots \texttt{A}}^{n} \to \texttt{A}
$$

It would be nice to get rid of repeated variables even if they don't appear together. Therefore, we develop the next lemma.

Lemma 4 $y \circ (x \circ y) \to x \circ y$

$$
\begin{array}{rl}
y \circ (x \circ y) & \text{Given} \\
y \circ (y \circ x) & \text{Commutativity} \\
(y \circ y) \circ x & \text{Associativity} \\
y \circ x & \text{Idempotence} \\
x \circ y & \text{Commutativity}
\end{array}
$$

`BAB` $\to$ `AB`

1.9   Connections to the Literature

The detailed example of the two variable Boolean function as presented in section 1.4, and later elaborated on in the exercises, was adapted from Frank Markham Brown's Example 3.7.1 as it appears in his book *Boolean Reasoning* [3]. I also borrowed the definitions for an n-variable function, a Boolean Algebra, Boole's Expansion Theorem, and the general tenor of a typical axiom system for a Boolean Algebra from the same source.

Brown uses the more conventional symbols of a Boolean Algebra like, $+, \cdot, 0, 1$ where I used $\bullet, \circ, F$, and T. As mentioned in the text, I wanted to emphasize the extreme abstractness of a Boolean Algebra.

Furthermore, Brown takes pains to emphasize, as others do not, that the Boolean part of Boolean Algebra doesn't refer to variables being able to assume just two values. That was why he and I thought it better to present a first example where the variables can take on four abstract values.

I found Brown's book *Boolean Reasoning: The Logic of Boolean Equations* to be fascinating and worthwhile. Eventually, he gets around to treating what he calls "syllogistic reasoning," but he never takes the path toward probability theory and inference.

It is curious thing that two human minds seem to be on a journey down the same road, but then, suddenly, one of the travelers takes a fork in the road, leaving the other to continue on alone down the other branch. Once separated, they hardly ever to seem to meet up again at the final destination that seemed so obvious at the outset.

Two other informative books on Boolean Algebra that I read for this Chapter are Hohn [11] and Schneeweiss [17].

For a wonderfully entertaining, and, at the same time, rigorous introduction to formal systems, read Chapters I and II of Douglas Hofstadter's, *Gödel, Escher, Bach* [10]. His explanation of theorems, axioms, and rules greatly influenced my style of presenting lemmas and theorems for the formal system known as Boolean Algebra.

1.10   Solved Exercises for Chapter One

Exercise 1.10.1: These beginning exercises illustrate the definition of a function as given on page 5. Illustrate Definition 1.

Solution to Exercise 1.10.1

Let the set S consist of $\{\clubsuit, \diamondsuit\}$. Let the set T consist of $\{\spadesuit, \heartsuit\}$. The function f assigns $f(\clubsuit) = \spadesuit$ and $f(\diamondsuit) = \heartsuit$ where $\clubsuit$ and $\diamondsuit$ are the elements of S and $f(x)$ belongs to the set T.

Exercise 1.10.2: Illustrate Definitions 2 and 3.

Solution to Exercise 1.10.2

$S \times S$ is the set of ordered pairs $\{(\clubsuit, \clubsuit), (\clubsuit, \diamondsuit), (\diamondsuit, \clubsuit), (\diamondsuit, \diamondsuit)\}$. A function f such that $f : S \times S \to S$ would assign to every element x of $S \times S$ an element $f(x) \in S$. For example, $f[(\clubsuit, \diamondsuit)] = \diamondsuit$. This is a binary operation since the set $S \times S$ will always consist of elements with two items. We are permitted to use any symbol we like for the function symbol f, so use $\circ$ and write $\circ[(\clubsuit, \diamondsuit)] = \diamondsuit$. This way of writing the function employs a prefix notation. If we want to use an infix notation, write instead $\clubsuit \circ \diamondsuit = \diamondsuit$.

Exercise 1.10.3: Illustrate Definitions 4 and 5.

Solution to Exercise 1.10.3

Let the set $\mathbf{B}$ consist of $\{\clubsuit, \diamondsuit, \spadesuit, \heartsuit\}$. The Cartesian Product $\mathbf{B} \times \mathbf{B}$ or $\mathbf{B}^2$ consists of all 16 ordered pairs $\{(\clubsuit, \clubsuit), (\clubsuit, \diamondsuit) \cdots (\heartsuit, \heartsuit)\}$. A binary operator $\bullet$ is a function f such that $f : \mathbf{B}^2 \to \mathbf{B}$. Thus, one example of f might be $f[(\spadesuit, \heartsuit)] = \clubsuit$. Or, written in infix notation, $\spadesuit \bullet \heartsuit = \clubsuit$.

Exercise 1.10.4: Confirm a couple more rows from the Boolean functional assignment table for two variables in Table 1.3.

Solution to Exercise 1.10.4

Take row 5 where $x = a'$ and $y = a$. Repeating the Boolean formula,

$$f(x, y) = (a' \circ x) \bullet (a \circ y')$$

we find that,

$$f(x = a', y = a) \quad = \quad (a' \circ a') \bullet (a \circ a')$$

$$= \quad a' \bullet F$$

$$= \quad a'$$

Take row 14 where $x = T$ and $y = a'$. Repeating the Boolean formula,

$$f(x, y) = (a' \circ x) \bullet (a \circ y')$$

we find that,

$$f(x = T, y = a') \quad = \quad (a' \circ T) \bullet (a \circ (a')')$$

$$= \quad a' \bullet (a \circ a)$$

$$= \quad a' \bullet a$$

$$= \quad T$$

Recall that the result of any operation involving $\circ$ can be found by consulting the binary operator table in Table 1.1, and the result of any operation involving $\bullet$ by consulting the binary operator table in Table 1.2. In the last example where variable y took on the value of a' and the formula told us to take the complement of the variable y, then another **Complementation** axiom, $(a')' = a$, comes into play.

Exercise 1.10.5: How many rows would a functional assignment table with three variables have?

Solution to Exercise 1.10.5

Assuming the same carrier set **B** as in the Table 1.3 example, there would be $4^3 = 64$ rows. The first row would show the functional assignment $f(x, y, z)$ for the setting of $x = a$, $y = a$, and $z = a$. The 64th and last row would show the functional assignment $f(x, y, z)$ for the setting of $x = T$, $y = T$, and $z = T$.

Exercise 1.10.6: Construct a simple DNF expansion of a Boolean function of two variables by manipulating the coefficient functions.

Solution to Exercise 1.10.6

Repeating Boole's Expansion Theorem for two variables,

$f(x, y) =$

$$[f(T, T) \circ x \circ y] \bullet [f(T, F) \circ x \circ y'] \bullet [f(F, T) \circ x' \circ y] \bullet [f(F, F) \circ x' \circ y']$$

try to arrange things such that we obtain a simple expansion consisting of say $x \circ y$. If we were to assign,

$$f(T, T) = T$$

$$f(T, F) = F$$

$$f(F, T) = F$$

$$f(F, F) = F$$

to all four coefficients $f(\star, \star)$ appearing the expansion, then,

$$f(x, y) = [\, T \circ x \circ y \,] \bullet [\, F \circ x \circ y' \,] \bullet [\, F \circ x' \circ y \,] \bullet [\, F \circ x' \circ y' \,]$$

The final three terms all disappear because of the presence of F together with the $\circ$ operator. The first term is all that remains so that,

$$f(x, y) = (T \circ x \circ y) \bullet F \bullet F \bullet F = x \circ y$$

As we will see in the next chapter, this is indeed the DNF expansion for the AND operator in Classical Logic.

Exercise 1.10.7: Verify the Associativity and Distributivity axioms with another example.

Solution to Exercise 1.10.7

Is $(x \bullet y) \circ z = (z \circ x) \bullet (z \circ y)$ for a setting of $x = F$, $y = a$, and $z = T$? The left hand side works out to a,

$$(x \bullet y) \circ z = (F \bullet a) \circ T = a \circ T = a$$

while the right hand side also works out to a,

$$(z \circ x) \bullet (z \circ y) = (T \circ F) \bullet (T \circ a) = F \bullet a = a$$

Exercise 1.10.8: If a Boolean Algebra has a carrier set with six elements, construct a Boolean function of your choosing by looking at a DNF expansion.

Solution to Exercise 1.10.8

Suppose that the carrier set is labeled as follows,

$$\mathbf{B} = \{a, a', b, b', F, T\}$$

Suppose further, that of the eight coefficient functions $f(\star, \star, \star)$, two are,

$$f(T, T, T) = b \text{ and } f(T, F, F) = a'$$

and the rest are equal to F. Then, using Boole's Expansion Theorem, a Boolean canonical formula can be constructed for the function as,

$$f(x, y, z) = bxyz \bullet a'xy'z'$$

x can be "factored out" and when $x = T$,

$$T \circ ((byz) \bullet (a'y'z')) = byz \bullet a'y'z'$$

Exercise 1.10.9: Accomplish something similar for a four variable Boolean formula.

Solution to Exercise 1.10.9

Keeping the same carrier set as in the last exercise, suppose that $f(T, T, T, F) = b'$ and with the remaining 15 coefficient functions all equal to F,

$$f(w, x, y, z) = b'wxyz'$$

Exercise 1.10.10: Create a Boolean Algebra with abstract symbols.

Solution to Exercise 1.10.10

The quintuple for an arbitrary abstract Boolean Algebra might look something like,

$$(\mathbf{B}, \flat, \natural, \heartsuit, \spadesuit)$$

Suppose that the carrier set $\mathbf{B}$ consists of the elements,

$$\mathbf{B} = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}.$$

The special elements are $\heartsuit$ and $\spadesuit$ and the binary operators are $\flat$ and $\natural$. The complement of $\clubsuit$ is $\diamondsuit$, while the complement of $\heartsuit$ is $\spadesuit$.

Exercise 1.10.11: Write out the syntactically correct formula for the first binary operation on the first element and its complement.

Solution to Exercise 1.10.11

The first element of $\mathbf{B}$ is $\clubsuit$ and its complement is $\diamondsuit$. The first binary operation is $\flat$ and we are employing the infix notation so the operator appears between its two arguments, thus, $\clubsuit \flat \diamondsuit$.

Exercise 1.10.12: Write out a syntactically correct formula for the second binary operation on the special elements and the first two elements, followed by the first binary operation on these results.

Solution to Exercise 1.10.12

Notice the placement of the parentheses in the solution,

$$(\heartsuit \, \natural \, \spadesuit) \, \flat \, (\clubsuit \, \natural \, \diamondsuit)$$

Exercise 1.10.13: Show the operator table for $\natural$.

Solution to Exercise 1.10.13

The binary operator table for $\natural$ is shown as Table 1.6,

Table 1.6: *The definition of the binary operator "$\natural$" for a Boolean Algebra by an operator table.*

$\natural$	$\clubsuit$	$\diamondsuit$	$\heartsuit$	$\spadesuit$
$\clubsuit$	$\clubsuit$	$\spadesuit$	$\clubsuit$	$\spadesuit$
$\diamondsuit$	$\spadesuit$	$\diamondsuit$	$\diamondsuit$	$\spadesuit$
$\heartsuit$	$\clubsuit$	$\diamondsuit$	$\heartsuit$	$\spadesuit$
$\spadesuit$	$\spadesuit$	$\spadesuit$	$\spadesuit$	$\spadesuit$

Exercise 1.10.14: Based on the above table, what is $(\diamondsuit \, \natural \, \heartsuit) \, \natural \, (\clubsuit \, \natural \, \diamondsuit)$?

Solution to Exercise 1.10.14

$$\diamondsuit \, \natural \, \heartsuit \rightarrow \diamondsuit \qquad \clubsuit \, \natural \, \diamondsuit \rightarrow \spadesuit \qquad \diamondsuit \, \natural \, \spadesuit \rightarrow \spadesuit$$

Exercise 1.10.15: Show a typical functional assignment for this Boolean Algebra.

Solution to Exercise 1.10.15

Using the binary operator $\natural$ as the function,

$$\natural : \mathbf{B}^2 \rightarrow \mathbf{B}$$

an ordered pair from $\mathbf{B} \times \mathbf{B}$ might be $(\heartsuit, \diamondsuit)$ and the mapping induced by $\natural$ is then $\diamondsuit$. We write the function as $f[(\heartsuit, \diamondsuit)] = \diamondsuit$.

Exercise 1.10.16: What are the elements of the domain for B × B.

Solution to Exercise 1.10.16

There are 16 elements in the set,

$$\mathbf{B} \times \mathbf{B} = \mathbf{B}^2 = \{ \overbrace{(\clubsuit, \clubsuit)}^{1}, \overbrace{(\clubsuit, \diamondsuit)}^{2}, \cdots, \overbrace{(\heartsuit, \diamondsuit)}^{10}, \cdots, \overbrace{(\spadesuit, \spadesuit)}^{16} \}$$

Exercise 1.10.17: Suppose that the Boolean formula, $(\diamondsuit \natural x) \natural y'$, defines a two variable function. What is $f(x, y)$ for the third element in the domain?

Solution to Exercise 1.10.17

If we suppose that the ordering of the elements continues to follow Exercise 1.10.16, the particular assignments for the two variables in this problem are $x \to \clubsuit$ and $y \to \heartsuit$. Therefore,

$$
\begin{aligned}
f(x, y) &= (\diamondsuit \natural \clubsuit) \natural \heartsuit' \\
&= \spadesuit \natural \spadesuit \\
&= \spadesuit
\end{aligned}
$$

Exercise 1.10.18: Find the binary operator table for ♭.

Solution to Exercise 1.10.18

The binary operator table for ♭ is shown in Table 1.7 below.

Table 1.7: *The definition of the binary operator " ♭ " for a Boolean Algebra by an operation table.*

♭	♣	◇	♡	♠
♣	♣	♡	♡	♣
◇	♡	◇	♡	◇
♡	♡	♡	♡	♡
♠	♣	◇	♡	♠

Exercise 1.10.19: Find the DNF for the Boolean formula $(\diamondsuit' \flat x) \natural (\diamondsuit \flat y')$.

Solution to Exercise 1.10.19

Conduct a pattern matching exercise on the DNF expansion formula appropriate for two variables as shown in section 1.5. The element $\spadesuit$ is the analog to the element T, while the element $\heartsuit$ is the analog to the element F. The analog to the binary operator $\flat$ is $\circ$, while the analog to the binary operator $\natural$ is $\bullet$.

Following this template, the DNF expansion for any formula becomes,

$$[f(\spadesuit,\spadesuit) \flat x \flat y] \natural [f(\spadesuit,\heartsuit) \flat x \flat y'] \natural [f(\heartsuit,\spadesuit) \flat x' \flat y] \natural [f(\heartsuit,\heartsuit) \flat x' \flat y']$$

Next, we have to find the four coefficient functions, $f(\spadesuit,\spadesuit)$, $f(\spadesuit,\heartsuit)$, $f(\heartsuit,\spadesuit)$, and $f(\heartsuit,\heartsuit)$ by substituting both arguments into the formula as given in the statement of the problem. We will need to consult both binary operator tables as we grind through the calculation.

$$
\begin{aligned}
f(\spadesuit,\spadesuit) &= (\diamondsuit' \flat \spadesuit) \natural (\diamondsuit \flat \spadesuit') \\[4pt]
&= \clubsuit \natural \heartsuit \\[4pt]
&= \clubsuit \\[8pt]
f(\spadesuit,\heartsuit) &= (\diamondsuit' \flat \spadesuit) \natural (\diamondsuit \flat \heartsuit') \\[4pt]
&= \clubsuit \natural \diamondsuit \\[4pt]
&= \spadesuit \\[8pt]
f(\heartsuit,\spadesuit) &= (\diamondsuit' \flat \heartsuit) \natural (\diamondsuit \flat \spadesuit') \\[4pt]
&= \heartsuit \natural \heartsuit \\[4pt]
&= \heartsuit \\[8pt]
f(\heartsuit,\heartsuit) &= (\diamondsuit' \flat \heartsuit) \natural (\diamondsuit \flat \heartsuit') \\[4pt]
&= \heartsuit \natural \diamondsuit \\[4pt]
&= \diamondsuit
\end{aligned}
$$

Thus, the DNF works out to,

$$(\clubsuit xy) \natural (\spadesuit xy') \natural (\diamondsuit x'y')$$

where, as before, we condense the notation by writing the constants and variables together to indicate the $\flat$ operator.

Exercise 1.10.20: Justify the elements in the ♭ binary operator table by the axioms of Boolean Algebra.

Solution to Exercise 1.10.20

The diagonal elements are filled in by the **Idempotence axiom**, the ♡ and ♠ rows and columns by the **Special Elements axiom**, and the two remaining ♣ and ♢ cells by the **Complementation axiom** and the **Commutativity axiom**.

Exercise 1.10.21: Section 1.4 presented a Boolean functional assignment table for two variables. Now suppose that the formula for the Boolean functional assignment in Table 1.3 had been specified differently as,

$$f(x, y) = (a \bullet x') \circ (a' \bullet y)$$

What would $f(x, y)$ for Row 1 now look like?

Solution to Exercise 1.10.21

Row 1 says to make the assignment of $x = a$ and $y = a$. Therefore,

$$
\begin{aligned}
f(a, a) &= (a \bullet a') \circ (a' \bullet a) \\
&= T \circ T \\
&= T
\end{aligned}
$$

Exercise 1.10.22: What would $f(x, y)$ for Row 8 evaluate to?

Solution to Exercise 1.10.22

Row 8 says to make the assignment of $x = a'$ and $y = T$. Therefore,

$$
\begin{aligned}
f(a', T) &= (a \bullet a) \circ (a' \bullet T) \\
&= a \circ T \\
&= a
\end{aligned}
$$

Exercise 1.10.23: Show that the original formula in section 1.5 and its DNF expansion are equivalent.

Solution to Exercise 1.10.23

In section 1.5, the function $f(x, y)$ was expanded into the canonical form,

$$f(x, y) = (a' \circ x \circ y) \bullet (x \circ y') \bullet (a \circ x' \circ y').$$

We would need to show that the DNF expansion results in the same values as the original formula for all sixteen possible variable settings as given in Table 1.3. As a start, look at row 7 where $x = a'$ and $y = F$. The functional assignment according to the original formula is $f(x, y) = T$. Substituting these variable assignments into the DNF formula,

$$
\begin{aligned}
(a' \circ x \circ y) \bullet (x \circ y') \bullet (a \circ x' \circ y') &= ((a' \circ a') \circ F) \bullet (a' \circ T) \bullet ((a \circ a) \circ T) \\
&= (a' \circ F) \bullet a' \bullet (a \circ T) \\
&= F \bullet (a' \bullet a) \\
&= T
\end{aligned}
$$

We have confirmation for one case. How can you prove that the two formulas are equivalent?

Exercise 1.10.24: Write out a *syntactically incorrect* Boolean formula.

Solution to Exercise 1.10.24

One hopes that all the Boolean formulas presented are syntactically correct. For example, a formula such as $(x \circ y) \bullet (y' \circ x)$ is constructed according to the proper syntax.

The raw ingredients for constructing a Boolean formula consist of symbols like the following, " (",") ","$\circ$", "$\bullet$", "x", "y", and " $\prime$ ". Assembling these raw ingredients in any way that doesn't make sense is syntactically incorrect. After a few examples, it becomes very clear that it is child's play to immediately construct an infinite number of syntactically incorrect Boolean formulas. Here are a few:

$$)(\bullet)$$

$$(\circ x ($$

$$(y \bullet' x)$$

Exercise 1.10.25: Explain how the template for the two variable DNF expansion is formed.

Solution to Exercise 1.10.25

This problem was inspired by Brown's solution for the so–called *minterm canonical form* [3], page 45. Brown credits Boole for the basic expansion theorem, which, when applied in a recursive manner, results in the DNF template for any number of variables.

In our notation, an n-variable Boolean function can be expanded quite generally as,

$$f(x_1, x_2, \cdots, x_n) = [\, x_1 \circ f(T, x_2, \cdots, x_n)\,] \bullet [\, x_1' \circ f(F, x_2, \cdots, x_n)\,]$$

Take the specific case of two variables, x and y, where, following the general template just given,

$$f(x, y) = [\, x \circ f(T, y)\,] \bullet [\, x' \circ f(F, y)\,]$$

Now apply Boole's Expansion Theorem in a recursive fashion to expand the new functions $f(T, y)$ and $f(F, y)$. Any computer program employing recursion keeps building a stack for the recursion to work on until some base condition is reached. Here the base condition is to reduce functions with any argument still a variable to functions with arguments involving only T or F.

$$f(T, y) = [\, f(T, T) \circ y\,] \bullet [\, f(T, F) \circ y'\,]$$

Substitute this expansion into the appropriate place.

$$x \circ f(T, y) = x \circ [\, f(T, T)\, y \bullet f(T, F)\, y'\,]$$

Use the **Distributivity axiom** to distribute the x. Then follow this operation with the **Commutativity axiom**,

$$x \circ f(T, y) = f(T, T)\, xy \bullet f(T, F)\, xy'$$

And in the same recursive way, expand $f(F, y)$ to complete the derivation of the template for the expansion of $f(x, y)$,

$$
\begin{aligned}
f(F, y) &= [\, f(F, T) \circ y\,] \bullet [\, f(F, F) \circ y'\,] \\[2mm]
x' \circ f(F, y) &= x' \circ [\, f(F, T)\, y \bullet f(F, F)\, y'\,] \\[2mm]
x' \circ f(F, y) &= f(F, T)\, x'y \bullet f(F, F)\, x'y' \\[2mm]
f(x, y) &= f(T, T)\, xy \bullet f(T, F)\, xy' \bullet f(F, T)\, x'y \bullet f(F, F)\, x'y'
\end{aligned}
$$

Exercise 1.10.26: Continue on in the same manner as section 1.8 and prove some more lemmas and theorems.

Solution to Exercise 1.10.26

Theorem 1 *Any number of $\circ$ operations on two variables in succession can be reduced and reordered without repetitions.*

Here is an example of such a theorem,

$$((y \circ y) \circ y) \circ ((((x \circ x) \circ x) \circ x) \circ x) \circ ((((y \circ y) \circ y) \circ y) \circ y) \qquad \text{Given}$$

$$y \circ ((((x \circ x) \circ x) \circ x) \circ x) \circ ((((y \circ y) \circ y) \circ y) \circ y) \qquad \text{Lemma 3}$$

$$y \circ (x \circ ((((y \circ y) \circ y) \circ y) \circ y)) \qquad \text{Lemma 3}$$

$$y \circ (x \circ y) \qquad \text{Lemma 3}$$

$$x \circ y \qquad \text{Lemma 4}$$

This theorem is more easily understood by its patterned representation,

$$\overbrace{\texttt{BBB}\cdots\texttt{B}}^{n_1} \overbrace{\texttt{AAA}\cdots\texttt{A}}^{n_2} \overbrace{\texttt{BBB}\cdots\texttt{B}}^{n_3} \to \texttt{AB}$$

It would be nice if a similar theorem were true for any number of variables. To that end, we develop another lemma,

Lemma 5 $z \circ (y \circ x) \to x \circ (y \circ z)$

$$z \circ (y \circ x) \qquad \text{Given}$$

$$z \circ (x \circ y) \qquad \text{Commutativity}$$

$$(z \circ x) \circ y \qquad \text{Associativity}$$

$$(x \circ z) \circ y \qquad \text{Commutativity}$$

$$x \circ (z \circ y) \qquad \text{Associativity}$$

$$x \circ (y \circ z) \qquad \text{Commutativity}$$

$\texttt{CBA} \to \texttt{ABC}$

Theorem 2 *Any number of $\circ$ operations on any number of variables can be reduced and reordered without repetitions.*

Here is an example of Theorem 2.

$$(((z \circ z) \circ z) \circ z) \circ (((y \circ y) \circ y) \circ y) \circ (((x \circ x) \circ x) \circ x)$$

$$z \circ (((y \circ y) \circ y) \circ y) \circ (((x \circ x) \circ x) \circ x) \qquad \text{Lemma 3}$$

$$z \circ y \circ (((x \circ x) \circ x) \circ x) \qquad \text{Lemma 3}$$

$$z \circ (y \circ x) \qquad \text{Lemma 3}$$

$$x \circ (y \circ z) \qquad \text{Lemma 5}$$

This theorem, just like the one above, is more easily understood by its patterned representation,

$$\overbrace{\texttt{CCC}\cdots\texttt{C}}^{n_1}\ \overbrace{\texttt{BBB}\cdots\texttt{B}}^{n_2}\ \overbrace{\texttt{AAA}\cdots\texttt{A}}^{n_3} \to \texttt{ABC}$$

Theorem 3 *The complement of a variable and a ∘ operation on an expression with the respective uncomplemented variable results in F.*

$$
\begin{array}{rl}
x' \circ ((z \circ (y \circ x))) & \text{Given} \\[1.5em]
x' \circ ((x \circ (y \circ z))) & \text{Lemma 5} \\[1.5em]
x' \circ ((x \circ y) \circ z) & \text{Associativity} \\[1.5em]
((x' \circ x) \circ y) \circ z & \text{Associativity} \\[1.5em]
(F \circ y) \circ z & \text{Complementation} \\[1.5em]
F \circ z & \text{Special Elements} \\[1.5em]
F & \text{Special Elements}
\end{array}
$$

The patterned representation looks like,

$$\overline{\texttt{A}}\texttt{CBA} \to \texttt{F}$$

Corollary: $y' \circ (x \circ y) \to F$

$$
\begin{array}{rl}
y' \circ (x \circ y) & \text{Given} \\[1.5em]
F & \text{Theorem 3}
\end{array}
$$

The patterned representation looks like,

$$\overline{\texttt{B}}\texttt{AB} \to \texttt{F}$$

Exercise 1.10.27: Suppose that the DNF for the Boolean formula with three variables is $(\clubsuit xyz) \, \natural \, (\diamondsuit xy'z)$. Using whatever lemmas or theorems that have been proven so far, find $f(\clubsuit, \clubsuit, \clubsuit)$.

Solution to Exercise 1.10.27

Lemma 2 was applied to the first term, and **Theorem 3** to the second term.

$$
\begin{aligned}
f(\clubsuit, \clubsuit, \clubsuit) &= (\clubsuit \; x \; y \; z) \, \natural \, (\diamondsuit \; x \; y' \; z) \\[2mm]
&= (((\clubsuit \, \flat \, \clubsuit) \, \flat \, \clubsuit) \, \flat \, \clubsuit) \, \natural \, (((\diamondsuit \, \flat \, \clubsuit) \, \flat \, \diamondsuit) \, \flat \, \clubsuit) \\[2mm]
&= \clubsuit \, \natural \, \heartsuit \\[2mm]
&= \clubsuit
\end{aligned}
$$

Exercise 1.10.28: Use the $\bullet$ operation to derive some more theorems.

Solution to Exercise 1.10.28

To this point, only the $\circ$ operation has been used to derive theorems. Here, we use the **Distributivity Axiom** to derive theorems involving the $\bullet$ operation. The next theorem is analogous to the traditional notion of "factoring out" a common factor. The following theorem helps us understand the disjunctive normal form a little better. It is analogous to orthonormal basis functions that sum to 1.

Theorem 4 $(x \circ y) \bullet (x' \circ y) \rightarrow y$

$$
\begin{aligned}
(x \circ y) \bullet (x' \circ y) &\quad \text{Given} \\[2mm]
(y \circ x) \bullet (y \circ x') &\quad \text{Commutativity (twice)} \\[2mm]
y \circ (x \bullet x') &\quad \text{Distributivity (in reverse)} \\[2mm]
y \circ T &\quad \text{Complementation} \\[2mm]
y &\quad \text{Special Elements}
\end{aligned}
$$

The patterned representation looks like,

$$
\texttt{AB} + \overline{\texttt{A}}\texttt{B} \rightarrow \texttt{B}(\texttt{A} + \overline{\texttt{A}}) \rightarrow \texttt{B}
$$

This analogy will be used later for the probability manipulation,

$$
P(AB \vee \overline{A}B) = P(B)
$$

Theorem 5 *The • operation on all the basis functions in the DNF results in T.*

Here the theorem is proved for two variables.

$$((x \circ y) \bullet (x' \circ y)) \bullet ((x \circ y') \bullet (x' \circ y')) \qquad \text{Given}$$

$$y \bullet ((x \circ y') \bullet (x' \circ y')) \qquad \text{Theorem 4}$$

$$y \bullet y' \qquad \text{Theorem 4}$$

$$T \qquad \text{Complementation}$$

The patterned representation looks like,

$$\mathtt{AB} + \overline{\mathtt{A}}\mathtt{B} + \mathtt{A}\overline{\mathtt{B}} + \overline{\mathtt{A}}\,\overline{\mathtt{B}} \to \mathtt{B}(\mathtt{A} + \overline{\mathtt{A}}) + \overline{\mathtt{B}}(\mathtt{A} + \overline{\mathtt{A}}) \to \mathtt{B} + \overline{\mathtt{B}} \to \mathtt{T}$$

and a later probability manipulation will derive,

$$P(AB) + P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) = 1$$

Theorem 6 (The Absorption axiom) $x \bullet (x \circ y) \to x$

This is another property of a Boolean Algebra which is sometimes given just as an axiom. It is known as the **Absorption axiom**. However, this seems a bit presumptuous given how many transformation rules have to be applied, so we go ahead and label it a theorem.

$$x \bullet (x \circ y) \qquad \text{Given}$$

$$(T \circ x) \bullet (x \circ y) \qquad \text{Special Elements (in reverse)}$$

$$(x \circ T) \bullet (x \circ y) \qquad \text{Commutativity}$$

$$x \circ (T \bullet y) \qquad \text{Distributivity (in reverse)}$$

$$x \circ T \qquad \text{Special Elements}$$

$$x \qquad \text{Special Elements}$$

Operating on x with T in the second step is like multiplying an expression by 1. The pattern looks like $\mathtt{A} + \mathtt{AB} \to \mathtt{A}$.

These abstract exercises with Boolean operations are important because the formal manipulation rules for probabilities are very similar. For example, we will see later that the absorption property works for the probability operation,

$$P(A \vee AB) = P(A)$$

mirroring Theorem 6.

Chapter 2

Logic Functions

2.1 Introduction

The conscious intent of the beginning Chapter was to keep things at an abstract level. No attempt was made to draw a connection between the formalities of Boolean Algebra and anything in the real world. But we depart from that pristine stance starting with this Chapter.

Classical Logic is one of the crowning achievements of human civilization. In essence, it attempts to tell us whether an argument to convince our fellow man makes any sense or not. Can you string together a number of agreed upon premises leading to a conclusion that must be accepted as correct? The ability to form an "air–tight" argument leading to the acceptance of an incontrovertible conclusion is the means whereby rational men try to change the course of history.

Classical Logic happens to be the next step to discovering how an Information Processor might conduct inferences and reason under uncertainty. Classical Logic is a *deductive* process that an Information Processor prefers to use under that happy circumstance when there is complete certainty. But more often than not, critical information is lacking, and the best an Information Processor can do is to make an *inference*.

We now take the leap that horrifies the pure mathematician. We dare to try and attach some *meaning* to the purely abstract nature of the mathematical structure as we have constructed it so far. Thus, Classical Logic will deal with ordinary statements that can be considered to be true or false.

Classical Logic is, in fact, just a special case of n-variable Boolean functions which we investigated in the first Chapter. It is actually a simpler scenario in that we need choose only two elements for the carrier set **B**. These two elements are just the two special elements, F and T, where F is the complement of T and vice versa. So it is fair to say that we going to examine the "algebra of logic."

Classical Logic is concerned with the case where every variable can assume just the values of F or T. The functional assignment to these variables must also come from F or T. Thus, Classical Logic is seen to be a specific case within the general definition of a Boolean function set out in the first Chapter as $f : \mathbf{B}^n \to \mathbf{B}$.

When, as is the case for Classical Logic, there are just two values for each variable, it becomes feasible to actually consider all possible functions for two and three variables. In contrast, there were over four billion possible functions for just two variables in the example of a Boolean Algebra with four elements as discussed in Chapter One.

Classical Logic will first be analyzed in the context of the 16 possible functions for two variables. Classical Logic attaches some familiar names to these functions such as OR and IMPLIES. These functions are studied by considering them as binary operators tying two variables together with some functional assignment. This was exactly how the abstract binary operators $\circ$ and $\bullet$ were presented in the beginning Chapter. Eventually, we will have to show how all 16 functions can be distilled down into functions of just three operators, AND, OR, and NOT.

Somewhat surprisingly, it is both conceptually and pragmatically easier to think about Classical Logic as these kinds of functional operations, rather than as some sort of mind–bending test of deductive prowess. In the course of demonstrating this, we will learn how to solve any formula from Classical Logic, no matter how complicated it might appear. And, of course, what we ultimately want to accomplish is to see how the formal aspects of probability theory generalize Classical Logic.

In the first example of Chapter One, we discussed a function as defined by its Boolean formula. We shall do the same thing here. The elements T and F lose some of their abstractness in this application because we are going to associate T with the meaning that a *statement* is TRUE, and F with the meaning that a *statement* is FALSE. The variables also lose some of their abstractness in logic applications because now we attach the meaning to variables that they are to represent these *statements* we have been talking about.

These statements are nothing more than the ordinary garden variety statements we face everyday. For example, here is a statement, "My first name is David." This statement is either TRUE or FALSE depending on the speaker. Let me propose another statement, "It will rain tomorrow." This statement is also either TRUE or FALSE depending on the details of how we might happen to define "rain" and after tomorrow has come and gone.

After an extensive discussion of the 16 functions defined on *two* variables, we will move up to all possible 256 functions defined on *three* variables at the end of the Chapter. The reason for doing this is to connect this slight extension of Classical Logic to elementary cellular automata.

Notice that we haven't yet mentioned anything about probability functions. Furthermore, we will, for the moment, retain that fascinating abstraction that no numbers shall make an appearance. Those developments will come later.

2.2   Functions and Formulas in Classical Logic

Following the formal definitions presented in the first Chapter, we write out the template for an $n = 2$ variable Boolean function as,

$$f : \mathbf{B}^2 \to \mathbf{B}$$

The carrier set $\mathbf{B}$ consists of only two elements, which are, at the same time, the special elements of the Boolean Algebra,

$$\mathbf{B} = \{\text{TRUE, FALSE}\}$$

The quintuple on which the Boolean Algebra is defined is thus,

$$(\mathbf{B}, \wedge, \vee, \text{TRUE, FALSE})$$

where we have introduced new symbols, the `AND` operator ($\wedge$) and the `OR` operator ($\vee$) as the two binary operators in this Boolean Algebra. For the sake of economy, we will continue to use the abstract symbols T and F to stand for the truth values TRUE and FALSE.

The $\wedge$ operator and the $\vee$ operator are seen to be analogous to the abstract binary operators $\circ$ and $\bullet$. And, just as before, these new binary operators are defined by a table. Tables 2.1 and 2.2 show the effects of the `AND` operator and the `OR` operator on the two elements of $\mathbf{B}$. In fact, these operators are two of the 16 possible Boolean functions.

Table 2.1: *The definition of the binary operator* `AND`*, the symbol* $\wedge$*, used in an "algebra of logic."*

$\wedge$	TRUE	FALSE
TRUE	TRUE	FALSE
FALSE	FALSE	FALSE

Table 2.2: *The definition of the binary operator* `OR`*, the symbol* $\vee$*, used in an "algebra of logic."*

$\vee$	TRUE	FALSE
TRUE	TRUE	TRUE
FALSE	TRUE	FALSE

In another notational change, instead of using variable names like x and y, we will switch to A and B for logic applications. Now, the interpretation of the symbols

is that A and B stand for statements, or propositions as they are sometimes called. These variables, now statements, can only assume the value of TRUE or FALSE.

So, keeping to our earlier example, A might be the statement, "My first name is David.", where A might be assigned the value of either TRUE or FALSE. B might be the statement, "It will rain tomorrow.", with once again, B assuming the value of either TRUE or FALSE.

The complement of any statement is now indicated by drawing a bar over the variable. So, the complement of A is written as $\overline{A}$ and indicates the statement, "My first name is NOT David" where negation will soon be defined as one of the 16 functions. We have examples hopefully defusing the potential confusion that might surround this tricky point.

The set $\mathbf{B}^2 \equiv \mathbf{B} \times \mathbf{B}$ is the set of all ordered pairs that could be constructed from $\mathbf{B}$. These can be listed explicitly as,

$\mathbf{B} \times \mathbf{B} = \{(\text{TRUE, TRUE}), (\text{TRUE, FALSE}), (\text{FALSE, TRUE}), (\text{FALSE, FALSE})\}$

and more compactly as,

$$\mathbf{B}^2 = \{\ (T,T), (T,F), (F,T), (F,F)\ \}$$

Therefore, one particular functional assignment $f : \mathbf{B}^2 \to \mathbf{B}$ where $A = \text{TRUE}$ and $B = \text{FALSE}$ might be,

$$f(A = \text{TRUE}, B = \text{FALSE}) = \text{FALSE}$$

which will be abbreviated to just $f(T, F) = F$.

The infix notation is also commonly used in Classical Logic as well as in Boolean Algebra. However, we will alternate between a notation like $f(T, F) = F$ and a notation like $T \wedge F = F$. The infix notation was the notational choice for the $\circ$ and $\bullet$ operators in Chapter One.

The first thing is to make sure that we know how to write out syntactically correct Boolean formulas using this new notation for Classical Logic. We wrote syntactically correct formulas like $x \circ y$, $y \bullet x$, and $(x \bullet y') \circ (x' \circ y)$ involving generic variables in an abstract Boolean Algebra. Equivalent syntactically correct formulas for Classical Logic look like $A \wedge B$, $B \vee A$, and $(A \vee \overline{B}) \wedge (\overline{A} \wedge B)$.

The syntactically correct expressions for the DNF expansions in Chapter One included terms like $(x \circ y) \bullet (x \circ y') \bullet (x' \circ y) \bullet (x' \circ y')$. Equivalent expressions in our current applications to Classical Logic would look like,

$$(A \wedge B) \vee (A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$$

Because we are defining Classical Logic as a Boolean Algebra, it must also obey all the properties of a Boolean Algebra. Therefore, we could write down a logical equivalency between two Boolean formulas,

$$A \vee B = B \vee A$$

as an example of the Commutativity property. In the generic Boolean Algebra, we wrote the **Commutativity axiom** as,

$$x \bullet y = y \bullet x$$

As an example of the **Distributivity axiom** we would write,

$$A \wedge (B \vee \overline{B}) = (A \wedge B) \vee (A \wedge \overline{B})$$

analogous to,

$$x \circ (y \bullet y') = (x \circ y) \bullet (x \circ y')$$

From the discussion in Chapter One, it might be expected that all Boolean formulas in logic could be written down as combinations of just two binary operators. It turns out that this is correct, but first we have to take a rather circuitous route to this conclusion by examining all possible functions that could be defined in the algebra of logic.

Then we will see that, fortunately, we don't need all 16 functions, but in fact just two suffice, namely the AND and OR operators along with a NOT operator analogous to the notion of the complement of an element. This is exactly what we did in the examples of the general Boolean Algebra in Chapter One. There we saw that the expansion of any given function to a canonical form, the disjunctive normal form, was quite helpful. The same kind of expansion will be applied to logic functions to show how the functions depend on the AND and OR operators.

2.3   The 16 Possible Functions

In Chapter One, we showed a function table, Table 1.3, listing all possible settings for two variables together with the actual functional assignment according to a given formula. This table explicitly defined just one of the huge number of possible functions. Since each of these two variables could take on one of four values, there were sixteen rows in this table, arising from $2^4 = 16$.

We now present the analogous tables for two logic variables, beginning with Table 2.3 at the top of the next page, with the variables labeled as A and B instead of x and y. We observe the new feature that only two values are possible for each variable instead of four in the previous case.

Table 2.4 does the same thing for the second function, $f_2(A, B)$. To generate the functional assignments in some systematic way, the last F in the functional assignment column from Table 2.3 was changed to a T. This leads to the new functional assignments as shown in the final column.

Thus, there are only four rows in these functional tables instead of sixteen rows, arising from $2^2 = 4$. Where before over four billion functions existed, there are now only sixteen. Therefore, it is conceivable to list and study them individually.

Table 2.3: *The first Boolean functional assignment table for two variables illustrating the Classical Logic functions.*

A	B	$f_1(A, B)$
T	T	F
T	F	F
F	T	F
F	F	F

Table 2.4: *A second Boolean functional assignment table illustrating the Classical Logic functions for two variables.*

A	B	$f_2(A, B)$
T	T	F
T	F	F
F	T	F
F	F	T

Table 2.3 presents the first function, $f_1(A, B)$, and Table 2.4 the second function from the totality of the 16 possible functional assignments. We could continue on in the same manner and generate all sixteen function tables.

However, notice that the first two columns for all such tables are going to be the same. For economy of presentation, we can shortcut this laborious process and present all 16 functions at once in one table.

Table 2.5 shows this rearrangement with the four possible values of the two variables as the columns. Each row defines the functional assignment for each one of the four particular settings of the variables. The next–to–last column presents a somewhat arbitrary assignment of a binary operator symbol to each one of these functions, while the final column presents some of the traditional names usually given to these functions.

Notice that the first two rows repeat the first two functions $f_1(A, B)$ and $f_2(A, B)$ discussed above, and that the two special binary operators of $\wedge$ and $\vee$ appear as functions $f_5(A, B)$ and $f_{15}(A, B)$, respectively.

Table 2.5: *All 16 functional assignments for two variables illustrating the Classical Logic functions.*

j	TT	TF	FT	FF	Symbol	$f_j(A, B)$
1	F	F	F	F	$\perp$	FALSE
2	F	F	F	T	$\downarrow$	NOR
3	F	F	T	F	$\star$	DIFFERENCE
4	F	T	F	F	$\diamond$	DIFFERENCE
5	T	F	F	F	$\wedge$	AND
6	F	F	T	T	$\vdash$	NOT A
7	F	T	F	T	$\dashv$	NOT B
8	T	F	F	T	$\leftrightarrow$	EQUAL
9	F	T	T	F	$\oplus$	XOR
10	T	F	T	F	$\triangleright$	B
11	T	T	F	F	$\triangleleft$	A
12	F	T	T	T	$\uparrow$	NAND
13	T	F	T	T	$\rightarrow$	IMPLIES
14	T	T	F	T	$\leftarrow$	IMPLIES
15	T	T	T	F	$\vee$	OR
16	T	T	T	T	$\top$	TRUE

2.4   Constructing Complicated Logic Formulas

In Chapter One, we wrote out syntactically correct Boolean formulas like,

$$f(x, y) = (a' \circ x) \bullet (a \circ y')$$

In this Chapter concerned with Classical Logic, we construct analogous Boolean formulas, which might, in the current notation, look something like this,

$$G(A, B) = (F \wedge A) \vee (T \wedge \overline{B})$$

We can build up complicated formulas in a recursive manner by selecting any of the binary operator symbols appearing in Table 2.5 to tie variables together. Parentheses are inserted to resolve any ambiguity. Here are a few examples,

$$F(A, B) \quad = \quad (A \perp B) \top A$$

$$G(A, B) \quad = \quad (A \star B) \oplus (A \to B)$$

$$H(A, B) \quad = \quad ((A \diamond B) \triangleright (A \leftarrow B)) \downarrow ((A \vdash B) \triangleleft (A \dashv B))$$

These formulas were indeed generated without any thought or purpose merely by inserting between A and B arbitrarily chosen binary operators from the 16 available. Then, think recursively as each new element is formed. The important thing is that we can form arbitrarily long Boolean formulas like these and they could be considered as Classical Logic formulas. These arbitrarily long formulas certainly look intimidating, but they can be mechanically generated at will.

Logic is usually introduced with just a few functions from the total of the 16 available functions rather than with these bizarre looking formulas. Function f_{13}, the "implication" operator, together with function f_8, the "if and only if" operator, are typically highlighted along with the "and" and "or" operators. Therefore, one is more likely to see formulas that look like,

$$F(A, B) = (A \vee B) \leftrightarrow (B \vee A)$$

or,

$$G(A, B) = (A \wedge (A \to B)) \to B$$

or,

$$H(A, B) = (\overline{A} \to B) \wedge (\overline{A} \to \overline{B}) \to A$$

After digesting the long and bizarre looking formulas, these are relatively benign. The point is simply that all of them are valid formulas using the defined binary operators.

Any of these formulas are more transparent and their solutions mechanically, albeit laboriously worked out, when written in a functional notation. All we have

to do is consult the function table, Table 2.5, for the appropriate operator's value for the particular setting of the variables.

We will illustrate this by solving each of the three "weird" formulas for some specific setting of the variables. This is another example of pushing abstract symbols around by formal rules as introduced in Chapter One. However, the game takes place with the new notation appropriate for Classical Logic, together with the barest hint of "meaning" that was absent from Boolean Algebra. Namely, we are talking about statements with truth values.

Solve the first formula for the particular settings of the two variables where $A = T$ and $B = T$. Work on the expression in parentheses first.

$$F(A, B) \;=\; (A \perp B) \top A$$

$$A \perp B \;=\; f_1(T, T)$$

$$f_1(T, T) \;=\; F$$

$$(A \perp B) \top A \;\equiv\; f_{16}\,[\,f_1(T, T),\; T\,]$$

$$f_{16}(F, T) \;=\; T$$

$$(A \perp B) \top A \;\equiv\; T$$

Now solve the second formula with a different variable assignment of $A = T$ and $B = F$. Work on the expressions within the parentheses first and then work on the $\oplus$ operator last.

$$G(A, B) \;=\; (A \star B) \oplus (A \to B)$$

$$A \star B \;=\; f_3(T, F)$$

$$=\; F$$

$$A \to B \;=\; f_{13}(T, F)$$

$$f_{13}(T, F) \;=\; F$$

$$(A \star B) \oplus (A \to B) \;=\; f_9\,[\,f_3(T, F),\; f_{13}(T, F)\,]$$

$$f_9(F, F) \;=\; F$$

The third formula, despite its complicated appearance, is solved in exactly the same way. Change the two variables to a setting of $A = F$ and $B = T$. Adopt a strategy of working from the inside out, and the solution, worked out in an exercise,

is shown to equal F,

$$((A \diamond B) \triangleright (A \leftarrow B)) \downarrow ((A \vdash B) \triangleleft (A \dashv B)) = F$$

2.5   Logical Tautologies

Logical tautologies might initially be construed as mind–bending logical puzzles. But if treated simply as the solution to some functional equations, they are worked out as mechanically as the previous exercises.

In this section, we begin to examine some logical tautologies that involve only two variables. After discussing three variable logic functions, we will then look at some tautologies involving three variables.

The point of a tautology is that it shows what the concept of *equal* must mean in logic. And, more importantly, illustrating that a tautology exists is the same as proving a theorem. The flip side of the coin to tautologies are logical contradictions.

Two forms, call them $\mathbf{F_1}$ and $\mathbf{F_2}$, are said to be *logically equivalent* if and only if $\mathbf{F_1} \leftrightarrow \mathbf{F_2}$ is a tautology, or, in other words, if this expression works out to T for all possible variable settings. We know how to do this from the examples of functional manipulation given in the previous section.

Let $\mathbf{F_1} \equiv A \vee B$ and $\mathbf{F_2} \equiv B \vee A$. Then if,

$$(A \vee B) \leftrightarrow (B \vee A)$$

has the value T for all four variable assignments, then $A \vee B$ is logically equivalent to $B \vee A$. This must be so if **Commutativity** is to hold for logic.

Systematically examine all four possible assignments to the variables A and B. For the first possible variable assignment of $A = T$ and $B = T$,

$$(A \vee B) \leftrightarrow (B \vee A) \quad = \quad f_8 \, [\, f_{15}(T,T), \; f_{15}(T,T) \,]$$

$$f_{15}(T,T) \quad = \quad T$$

$$f_8(T,T) \quad = \quad T$$

For the second possible variable assignment of $A = T$ and $B = F$,

$$(A \vee B) \leftrightarrow (B \vee A) \quad = \quad f_8 \, [\, f_{15}(T,F), \; f_{15}(F,T) \,]$$

$$f_{15}(T,F) \quad = \quad T$$

$$f_{15}(F,T) \quad = \quad T$$

$$f_8(T,T) \quad = \quad T$$

In like manner, the final two possible variable assignments would also work out to T. Thus, we have demonstrated[1] that a tautology exists. If there is a tautology, then we can correctly assert that $A \vee B$ is logically equivalent to $B \vee A$.

One must be careful and proceed slowly in constructing these tautologies. For example, one explanation for `EQUAL` is that this operator represents an implication in both directions (hence the symbology for the operator notation). Synonyms for this operator are `BICONDITIONAL` or `EQUIVALENT`. If $A \to B$, then the `BICONDITIONAL` operator means that $B \to A$ as well. Translating this formally we have,

$$\mathbf{F_1} \equiv (A \to B) \wedge (B \to A)$$

Then the `BICONDITIONAL` operator

$$\mathbf{F_2} \equiv A \leftrightarrow B$$

must be logically equivalent to $\mathbf{F_1}$. In other words,

$$((A \to B) \wedge (B \to A)) \leftrightarrow (A \leftrightarrow B)$$

For a change of pace, let's work out this example where the variable assignment is the third possible one, $A = F$ and $B = T$. In function notation, this potential tautology gets translated into a nested set of functions,

$$f_8 \{ f_5 [f_{13}(F,T), f_{13}(T,F)], f_8(F,T) \}$$

Referring back to Table 2.5 to find the functional assignment for the given variable settings, and working from the inside out, we have,

$$f_8(F,T) = F$$

$$f_{13}(F,T) = T$$

$$f_{13}(T,F) = F$$

$$f_5(T,F) = F$$

$$f_8(F,F) = T$$

The other three variable assignments also work out to T, so the above expression is a logical tautology. It is a proved theorem (all such theorems sound stilted when expressed verbally) that B is true if and only if A is true is the same as A implies B and B implies A.

If a Boolean expression evaluates to F for all possible variable assignments, then the expression is a contradiction. Here is an example of a contradiction. Note the use of parentheses indicating the order in which the operations are to be carried out.

$$(A \vee B) \wedge (\overline{A} \wedge \overline{B})$$

---

[1]Eventually, we will use *Mathematica* to alleviate the tedium of proving tautologies like these, but first we have to solve them by hand.

Examine the case where the particular variable assignment is $A = T$ and $B = T$. In the functional notation, this expression is then,

$$f_5 \{ f_{15}(T,T), f_5 [f_6(T,T), f_7(T,T)] \}$$

Substituting from the inside out, we have,

$$f_6(T,T) \quad = \quad F$$

$$f_7(T,T) \quad = \quad F$$

$$f_5(F,F) \quad = \quad F$$

$$f_{15}(T,T) \quad = \quad T$$

$$f_5(T,F) \quad = \quad F$$

And, of course, one would have to verify that the other three possible variable assignments also equal F.

Therefore, as a statement, the above Boolean expression is certain to be FALSE. In this case, one can reason that either A or B, or both, are TRUE, but at the same time, both A and B are FALSE. So how could it ever happen that either A or B or both are TRUE? It cannot happen under the circumstances just related by this expression. Therefore, the expression always returns FALSE because it reflects a contradiction.

2.6   Formulas from a Restricted Set of Functions

Now things would be pretty complicated if we had to deal with these 16 functions all the time. It would be even worse when we moved up to functions of three, four, or more variables. Do we now have to define 256 functions, 65,536 functions, and so on? It obviously would be a hopeless enterprise to do any kind of logical reasoning under these circumstances.

Fortunately, all these potential functions can be expressed in terms of a much smaller set of functions. For example, we will list some of the so–called *canonical* or *disjunctive normal forms* that employ just the $\wedge$ and $\vee$ operators together with the $\vdash$ and $\dashv$ operators represented by $\overline{A}$ and $\overline{B}$. These normal forms are *logically equivalent* to the operators listed in Table 2.5.

Table 2.6 at the top of the next page is constructed just like Table 2.5, but with a new final column that shows the canonical form defined in terms of just three operators instead of distinct operators for each function. The first and last functions could have been expressed in full DNF format, but they reduce to F and T. Some of the other functions can also be simplified from their full DNF expansion.

Table 2.6: *All 16 possible logic functions for two variables expressed in a canonical form using just the $\wedge$, $\vee$, and overbar operators.*

j	TT	TF	FT	FF	f_j	Canonical form $f_j(A, B)$
1	F	F	F	F	$\perp$	F
2	F	F	F	T	$\downarrow$	$\overline{A} \wedge \overline{B}$
3	F	F	T	F	$\star$	$\overline{A} \wedge B$
4	F	T	F	F	$\diamond$	$A \wedge \overline{B}$
5	T	F	F	F	$\wedge$	$A \wedge B$
6	F	F	T	T	$\vdash$	$(\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$
7	F	T	F	T	$\dashv$	$(A \wedge \overline{B}) \vee (\overline{A} \wedge \overline{B})$
8	T	F	F	T	$\leftrightarrow$	$(A \wedge B) \vee (\overline{A} \wedge \overline{B})$
9	F	T	T	F	$\oplus$	$(A \wedge \overline{B}) \vee (\overline{A} \wedge B)$
10	T	F	T	F	$\triangleright$	$(A \wedge B) \vee (\overline{A} \wedge B)$
11	T	T	F	F	$\triangleleft$	$(A \wedge B) \vee (A \wedge \overline{B})$
12	F	T	T	T	$\uparrow$	$(A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$
13	T	F	T	T	$\rightarrow$	$(A \wedge B) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$
14	T	T	F	T	$\leftarrow$	$(A \wedge B) \vee (A \wedge \overline{B}) \vee (\overline{A} \wedge \overline{B})$
15	T	T	T	F	$\vee$	$(A \wedge B) \vee (A \wedge \overline{B}) \vee (\overline{A} \wedge B)$
16	T	T	T	T	$\top$	T

2.6.1 Canonical forms and logical equivalency

We are going to prove that these new canonical forms are logically equivalent to the operators for the functions as given previously. Thus in some general sense, we want to establish logical equivalence between an old and a new form.

We conveniently label the old operator form as $\mathbf{F_{Old}}$ and the new canonical form as $\mathbf{F_{New}}$. The two forms are logically equivalent if and only if $\mathbf{F_{Old}} \leftrightarrow \mathbf{F_{New}}$ is a tautology. In logic, as we observed in the last section, a tautology means that $\mathbf{F_{Old}} \leftrightarrow \mathbf{F_{New}}$ has the value T for every possible variable assignment.

As an initial example, consider the intuitively obvious proof of the equivalence of the two ways of writing f_5. We illustrate the method for the first possible assignment to the variables, $A = T$ and $B = T$. The formula for the tautology is,

$$(A \wedge B) \leftrightarrow (A \wedge B)$$

Applying the definitions in a nested manner,

$$
\begin{aligned}
(A \wedge B) \leftrightarrow (A \wedge B) &\equiv f_8 \left[\, f_5(T,T), \, f_5(T,T) \, \right] \\
&= f_8(T,T) \\
&= T
\end{aligned}
$$

It turns out, as expected, that $(A \wedge B) \leftrightarrow (A \wedge B)$ is also assigned the value T for the other three possible assignments to the variables. One can see this immediately because f_5 is always F for the other three assignments and $f_8(F, F)$ is T.

All the other equivalencies are worked in the same fashion, although there is much careful work to get to the desired tautologies. For example, let's establish the logical equivalency for the two ways of writing f_9.

$$\mathbf{F_{Old}} \equiv A \oplus B$$

and

$$\mathbf{F_{New}} \equiv (A \wedge \overline{B}) \vee (\overline{A} \wedge B)$$

To satisfy logical equivalency, the following expression must result in T for all four variable assignments,

$$A \oplus B \leftrightarrow ((A \wedge \overline{B}) \vee (\overline{A} \wedge B))$$

First, let's establish what the overbar notation means in terms of one the 16 functions.

$$\overline{A} \equiv A \vdash B \equiv f_6(A, B)$$

The full DNF expansion of $f_6(A, B)$ is $(\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$. Using the **Distributivity** axiom of Boolean Algebra in the reverse direction,

$$(\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B}) = \overline{A} \wedge (B \vee \overline{B})$$

Applying **Complementation** and **Special Elements**, Equations (1.13) and (1.3), this gets turned into,

$$B \vee \overline{B} = T \text{ and } \overline{A} \wedge T = \overline{A}$$

This is why the function was called NOT A. Similarly, $f_7(A, B)$ is NOT B,

$$\overline{B} \equiv A \dashv B \equiv f_7(A, B)$$

For the first variable assignment of $A = T$ and $B = T$, the last term on the right hand side of $\mathbf{F_{New}}$ then becomes,

$$\overline{A} \wedge B \equiv f_5 [f_6(T, T), T]$$

Likewise, the first term on the right hand side of $\mathbf{F_{New}}$ then becomes,

$$A \wedge \overline{B} \equiv f_5 [T, f_7(T, T)]$$

Next, applying the OR operation, function f_{15}, the entire right hand side of $\mathbf{F_{New}}$ becomes,

$$(A \wedge \overline{B}) \vee (\overline{A} \wedge B) \equiv f_{15} [f_5 [T, f_7(T, T)], f_5 [f_6(T, T), T]]$$

The XOR operation on the left hand side, $\mathbf{F_{Old}}$, is,

$$A \oplus B \equiv f_9(T, T)$$

and, finally, putting both the left hand and right hand sides $\mathbf{F_{Old}}$ and $\mathbf{F_{New}}$ under the `EQUAL` operator,

$$\mathbf{F_{Old}} \quad \leftrightarrow \quad \mathbf{F_{New}}$$

$$A \oplus B \quad \leftrightarrow \quad ((A \wedge \overline{B}) \vee (\overline{A} \wedge B))$$

$$\equiv \quad f_8 \{f_9(T,T), f_{15} [\, f_5 \, [\, T, f_7(T,T) \,] \,], f_5 \, [\, f_6(T,T), T \,] \,] \}$$

Substituting the functional assignments from Table 2.5 yields,

$$A \oplus B \leftrightarrow ((A \wedge \overline{B}) \vee (\overline{A} \wedge B)) \quad \equiv \quad f_8 \{f_9(T,T), f_{15} [\, f_5(T,F), f_5(F,T) \,] \}$$

$$\equiv \quad f_8 \, [\, f_9(T,T), f_{15}(F,F) \,]$$

$$\equiv \quad f_8(F,F)$$

$$\equiv \quad T$$

And, if we worked this formula out for the other three possible choices for the variables the results would also yield T. Therefore, we have shown that the `XOR` operator is logically equivalent to the disjunction of the two conjunction terms, $A \wedge \overline{B}$ and $\overline{A} \wedge B$.

2.6.2   Heuristic for expanding to full disjunctive normal form

The expressions in the final column of Table 2.6 are written in the disjunctive normal form and abbreviated with the acronym DNF. This is the same concept that was used to study Boolean formulas in Chapter One. There is a heuristic pattern that can be followed in order to generate the expressions in the final column of Table 2.6.

The four functions f_2 through f_5 take on the functional assignment of T at only one particular setting for the variables. For example, f_5 is T only at the variable settings of $A = T$ and $B = T$ and F at the other three possibilities. This is exactly what is meant by A `AND` $B \equiv A \wedge B$.

Take another example where f_2 is T only when $A = F$ and $B = F$ and F at the other three possibilities. This is exactly what is meant by A `NOR` $B \equiv \overline{A} \wedge \overline{B}$. $A \downarrow B$ is T only when neither A nor B are T.

The expressions for the remaining functions f_6 through f_{16} can be generated by looking at those variable settings where the functional assignment takes on the value T. Then, use the `OR` operator ($\vee$) to join those expressions from f_2 through f_5 corresponding to where this T occurs.

For example, f_6 takes on the value T for two possible settings of the variables A and B. Specifically, when variables A and B take on the settings of,

$$A = F, B = F \text{ and } A = F, B = T$$

f_2 and f_3 take on a value of T here as well. Therefore,

$$f_6 \equiv (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$$

this being a disjunction of the two terms defining f_3 and f_2.

For another example, f_{12} has the functional assignment of T at three of the four possible settings of its two variables. These correspond to f_4, f_3, and f_2 so,

$$f_{12} \equiv (A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$$

Boole's Expansion Theorem, translated over into the new notation for Classical Logic, results in a template analogous to the one developed in Chapter One when the DNF was first discussed for an abstract Boolean Algebra. For two logical variables, a function is expanded into the canonical form of the DNF by following,

$$f(A, B) = [\, f(T,T)\, AB \,] \vee [\, f(T,F)\, A\overline{B} \,] \vee [\, f(F,T)\, \overline{A}B \,] \vee [\, f(F,F)\, \overline{A}\,\overline{B} \,]$$

where, for economy's sake, the $\wedge$ operator is implicitly assumed inside the brackets.

Since the functional assignments in Classical Logic must either be T or F, the DNF template simplifies. Whenever $f(\star, \star) = T$, that term is included in the DNF expansion; likewise, whenever $f(\star, \star) = F$, that term is not included in the DNF expansion. Hence, we see how the number of terms in the canonical forms of Table 2.6 arise. For example, the DNF expansion for $f_{15}(A, B)$, the OR operator, consists of three terms because,

$$
\begin{aligned}
f(T,T) &= T \\
f(T,F) &= T \\
f(F,T) &= T \\
f(F,F) &= F
\end{aligned}
$$

Thus, AB, $A\overline{B}$, and $\overline{A}B$ are included in the expansion, while $\overline{A}\,\overline{B}$ is excluded.

2.7   A Tautology for Three Variables

Is the logic expression, $A \rightarrow (B \rightarrow C)$, consisting of three variables, logically equivalent to the logic expression $(A \wedge B) \rightarrow C$? If we can show that,

$$(A \rightarrow (B \rightarrow C)) \leftrightarrow ((A \wedge B) \rightarrow C)$$

is T for all *eight* possible settings of the three variables, that is, if it is a tautology, then these two expressions are logically equivalent.

Like the previously worked examples, we will show that the above expression does, in fact, work out to T for one possible setting of the variables. Take one of the eight possible settings of the variables to be $A = T$, $B = F$, and $C = T$. Even though there are now three variables, this doesn't affect the fact that all operations are still binary operations. That is, we can still write out the functional format recursively with any of the sixteen functions taking on just two arguments.

Thus, the functional formula for the logical equivalence of the two expressions is written as,

$$f_8 \{ f_{13} [T, f_{13}(F, T)], f_{13} [f_5(T, F), T] \}$$

where the specific variable settings are filled in. As before, f_5 is the AND operator, f_{13} is the IMPLIES operator, and f_8 is the EQUAL operator. Working from the inside out, we find that this particular variable setting does result in T.

As we advance to functions with more variables, it is imperative to have a computer program calculate these tautologies. Suffice it to say that, as illustrated in Appendix A, a *Mathematica* program checked all eight possible settings of the three variables in the above formula and all possibilities returned the value T.

2.8 Three Variable Logic Functions

The Classical Logic functions on two variables can be extended to functions on three variables. The same Boolean definition for a function can be used as a template for more than two variables. Let's look in detail at three variable functions. The general abstract definition for a Boolean function of three variables now applies.

$$f : \mathbf{B}^3 \to \mathbf{B}$$

Thus, we will be writing functions on the ordered triples of $\mathbf{B} \times \mathbf{B} \times \mathbf{B}$ that take on values from $\mathbf{B}$, such as,

$$f(T, T, T) = T \text{ and } f(T, F, T) = F$$

There are 2^{2^n} possible functions for n variables which can each take on only one of two values. Previously, we have discussed the $2^{2^2} = 16$ possible functions for $n = 2$ variables. Now we are going to advance to the $2^{2^3} = 256$ functions, $f_1(A, B, C)$ through $f_{256}(A, B, C)$, for $n = 3$ variables.

No special names will be attached to any of these 256 functions on three variables as was done for the 16 functions over two variables. Also, we will want to eventually reduce every one of these functions into one of its canonical forms, the disjunctive normal form, just as we did for the two variable functions. We will focus, however, on one particular three variable logic function. This is the logic function underlying Wolfram's Rule 110 elementary cellular automaton.

Table 2.7 below shows the first possible function $f_1(A, B, C)$ which, analogously to $f_1(A, B)$, takes on the constant value of F for every one of the eight possible variable settings.

Table 2.7: *A Boolean functional assignment table for three variables illustrating the first of the 256 possible logic functions.*

A	B	C	$f_1(A, B, C)$
T	T	T	F
T	T	F	F
T	F	T	F
T	F	F	F
F	T	T	F
F	T	F	F
F	F	T	F
F	F	F	F

Table 2.8 displays this first function and the next eight functions $f_2(A, B, C)$ through $f_9(A, B, C)$ which assume the value T at just one of the possible eight settings. These are the analogous functions to the four functions $f_2(A, B)$ through $f_5(A, B)$ for the two variable case. As usual, for notational economy the $\wedge$ operator is suppressed. For example, the last column for f_2 should read $\overline{A} \wedge \overline{B} \wedge \overline{C}$.

Table 2.8: *The first function from above together with the next eight functions of three variables which take on the value T at only one particular setting of the variables. The disjunctive normal form is shown in the final column.*

j	TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF	f_j
1	F	F	F	F	F	F	F	F	F
2	F	F	F	F	F	F	F	T	$\overline{A}\,\overline{B}\,\overline{C}$
3	F	F	F	F	F	F	T	F	$\overline{A}\,\overline{B}C$
4	F	F	F	F	F	T	F	F	$\overline{A}B\overline{C}$
5	F	F	F	F	T	F	F	F	$\overline{A}BC$
6	F	F	F	T	F	F	F	F	$A\overline{B}\,\overline{C}$
7	F	F	T	F	F	F	F	F	$A\overline{B}C$
8	F	T	F	F	F	F	F	F	$AB\overline{C}$
9	T	F	F	F	F	F	F	F	ABC

Using Table 2.8 as a guide, all of the remaining functions can be constructed as disjunctions of these primitive expressions. For example, the functional assignment table for $f_{10}(A, B, C)$ looks like Table 2.9. A Boolean formula for this function is

Table 2.9: *A Boolean functional assignment table for three variables illustrating the tenth of the 256 possible logic functions.*

A	B	C	$f_{10}(A, B, C)$
T	T	T	F
T	T	F	F
T	F	T	F
T	F	F	F
F	T	T	F
F	T	F	F
F	F	T	T
F	F	F	T

constructed from the DNF expansion as,

$$f_{10}(A, B, C) \equiv (\overline{A}\,\overline{B}\,\overline{C}) \vee (\overline{A}\,\overline{B}C)$$

Notice that this is the disjunction of the two expressions for f_2 and f_3 shown in Table 2.8, the two functions where a T is also assigned for the same variable settings.

As a final example of the canonical form for a three variable logic function, consider the following function which will play the pivotal rule in the next Chapter when we take up cellular automata. The function consists of the disjunction of five terms where the assignment is T.

Table 2.10 shows all eight possible ordered triples from $\mathbf{B} \times \mathbf{B} \times \mathbf{B}$ with the corresponding functional assignment from $\mathbf{B}$. These represent the second, third, fifth, sixth, and seventh columns of Table 2.8. Therefore, following the pattern substitution rule, the disjunctive normal form is,

Table 2.10: *A Boolean functional assignment table for three variables illustrating one of the 256 possible logic functions. This function is used as the rule for Wolfram's Rule 110 cellular automaton.*

A	B	C	$f_{\star}(A, B, C)$
T	T	T	F
T	T	F	T
T	F	T	T
T	F	F	F
F	T	T	T
F	T	F	T
F	F	T	T
F	F	F	F

$$f_\star(A, B, C) = AB\overline{C} \vee A\overline{B}C \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$$

Once again, we emphasize that the function has been expanded so that only the three binary operations of `AND`, `OR`, and `NOT` need be consulted.

2.9   Connections to the Literature

A good, succinct, and to the point overview of the formal properties of Classical Logic that I found helpful can be found in Chapter 1 of Mendelson's *Boolean Algebra and Switching Circuits* [16]. He titles his introductory Chapter "The Algebra of Logic," an apt phrase I have borrowed because of its emphasis on a mechanical–like methodology for formally manipulating logic expressions.

Mendelson presents and solves typical logical word puzzles. I repeat one of them here because most people are initially exposed to logic by trying their hand at solving such puzzles. The earnest advice is to "reason logically." This kind of introduction discourages most people from pursuing the matter any further because these problems are invariably difficult to solve even though the most strenuous and diligent application of the solver's reasoning prowess has been called upon.

I intend to show that probability theory, to be considered as a generalization of Classical Logic, can indeed provide a simplifying framework for attacking such "logic problems." Chapter Ten delves a little bit more into "logical brain–twisters" after we have built up some preliminary concepts in probability necessary for handling them from the inferential viewpoint.

Here is Mendelson's "brain–twister," [16], pg. 25, Problem 1.17(a).

> Either Arlen is lying or Brewster was in Mexico in April or Crawford was not a blackmailer. If Brewster was not in Mexico in April, then either Arlen is telling the truth or Crawford was a blackmailer. Hence Brewster must have been in Mexico in April.

Does the conclusion that Brewster must have been in Mexico in April follow "logically" from the stated premises? Can we claim to have presented a "logical argument" if it were to turn out that the initial assumptions are *logically equivalent* to the conclusion?

In this Chapter, we studied how to set up and mechanically test such tautologies. If a tautology does not exist, then there is no *logical* equivalency between the premises and the conclusion. On one side we have the assumptions, and on the other side the conclusion. Are the assumptions logically equivalent to the conclusions?

Or, in other words, if the logic binary operator of `EQUAL` inserted between the assumptions and the conclusion is TRUE for all possible settings of the variables, then we have equivalency and the conclusions do follow from the assumptions.

As mentioned, Classical Logic, even though it is a purely deductive system, removes some of the abstractness inherent in that other purely deductive system that is Boolean Algebra. Logic is concerned with sentences that have a "truth value." This logic problem involves three sentences:

1. A: "Arlen is telling the truth."

2. B: "Brewster was in Mexico in April."

3. C: "Crawford is a blackmailer."

So the first assumption, "Either Arlen is lying or Brewster was in Mexico in April or Crawford is not a blackmailer. " is written symbolically as the expression,

$$\overline{A} \vee B \vee \overline{C}$$

The second assumption, "If Brewster was not in Mexico in April, then either Arlen is telling the truth or Crawford is a blackmailer." is written symbolically as the expression,

$$\overline{B} \to (A \vee C)$$

Now, we join both of these assumptions by the `AND` operator to arrive at the first form, $\mathbf{F_1}$,

$$(\overline{A} \vee B \vee \overline{C}) \wedge (\overline{B} \to (A \vee C))$$

The second form, $\mathbf{F_2}$, is simply the conclusion, B.

Does a tautology exist between the first form and the second form? In other words, if we test all eight possible settings for the three variables, will they all result in TRUE for $\mathbf{F_1} \leftrightarrow \mathbf{F_2}$, or,

$$(\overline{A} \vee B \vee \overline{C}) \wedge (\overline{B} \to (A \vee C)) \leftrightarrow B?$$

The whole point of all those boring tautology demonstrations was to emphasize that the unwieldy expression above is no cause for alarm. Having set it up properly, it can be worked through slowly and methodically by applying each binary operator for two variables as they are called out in the expression. We could check the variables starting with $A = T, B = T, C = T$ and ending with $A = F, B = F, C = F$ to see whether TRUE emerges as the outcome in all cases.

What we would find is that at two of the possible eight variable settings, our potential tautology yields FALSE. The conclusion does not, in fact, follow logically from the two assumptions. We cannot logically conclude that Brewster was in Mexico in April for all possible situations.

The first variable setting where the tautology fails is at $A = T, B = F$ and $C = F$ and the other variable setting is at $A = F, B = F$ and $C = T$. I avoided all that hand labor by writing a small *Mathematica* program and transferred the burden of checking to machine intelligence.

Compare the outcome from this mechanical procedure to the mental and verbal gyrations we have to go through to convince ourselves, if we ever do, that the argument is not logical. From the use of the EQUAL operator involved in the very definition of logical equivalency, $\mathbf{F_1} \leftrightarrow \mathbf{F_2}$, if we ever have F for B as we do for the first variable setting just mentioned, then we better have F as well for $\mathbf{F_1}$ if we are going to end up with T as a result of the binary operation EQUAL. So,

$$\mathbf{F_1} \equiv (\overline{A} \vee B \vee \overline{C}) \wedge (\overline{B} \to (A \vee C))$$

must work out to F. But it doesn't. It works out to T.

The first term in parentheses is $(F \vee F \vee T) \equiv T$. The second term is a nested expression. The inner parentheses works out to $(T \vee F) \equiv T$ with then the outer parentheses reducing to $(T \to T) \equiv T$. Finally, we have $(T \wedge T) \equiv T$, leading to $\mathbf{F_1}$ having the value T and thus destroying any chance for the tautology to exist.

If those convolutions were not enough, imagine trying to reason it out to yourself as follows: We will save ourselves a lot of trouble by considering just one of the circumstances that we already know ruins the tautology. This is $A = T$, $B = F$, and $C = F$.

So the first assumption, "Either Arlen is lying, or Brewster was not in Mexico in April, or Crawford is not a blackmailer." is TRUE because at least one of them is TRUE. In fact, two of them are TRUE, "Brewster was not in Mexico in April." and "Crawford is not a blackmailer."

Now, the second assumption, "If Brewster was not in Mexico," which he was not, remember $B = F$, implies that "Arlen was telling the truth." or that, "Crawford is a blackmailer." Now, the first sentence, "Arlen was telling the truth." is TRUE, while the second sentence, "Crawford is a blackmailer." is FALSE. However, these two sentences are joined by the OR operator, thus only one sentence, the one about Arlen, has to be TRUE for the joint sentence to be TRUE. Thus, the implication is TRUE.

The first joint sentence and the second joint sentence are both TRUE because of the definition of the AND operator, leading to whole left hand side of the proposed tautology being TRUE. That is, all of the assumptions are TRUE. However, the conclusion "Brewster was in Mexico." is FALSE, remember that $B = F$. We have one side that evaluates to TRUE, and the other side that evaluates to FALSE, so both sides can not be considered logically equivalent.

Isn't it easy to reason logically?

I will demonstrate to you later on that if you embed such conundrums within the framework of probability theory, then the correct logical conclusions follow inexorably from an application of mechanical rules. More importantly, even if trying to "reason logically" fails you, probability theory as inference will provide you with at least some sort of quantitative answer because probability generalizes logical reasoning.

Personally, the main impetus for actually contemplating whether probability theory could generalize logic was Jaynes's insistence on this notion. He actually opens his *Probability Theory: The Logic of Science* [12] with a discussion of these matters.

Jaynes provides an excellent introduction to Boolean Algebra on pages 9–17 of his book. It was working through his examples that lent some structure to my own take on logic functions as introduced here in Chapter Two.

Furthermore, he illustrates how either the `NAND` or the `NOR` logic functions can serve as an adequate set of logic functions by themselves. For example, he lists how the operations of negation, conjunction, and disjunction, (`NOT, AND, OR`) can be represented by the `NAND` function [12, pg. 16].

> It came as a pleasant surprise to find that there is not only one but two such operations. The operation 'NAND' is defined as the negation of 'AND':
>
> $$A \uparrow B \equiv \overline{AB} = \overline{A} + \overline{B}$$
>
> which we can read as 'A NAND B'. But then we have at once

$$\overline{A} = A \uparrow A$$

$$AB = (A \uparrow B) \uparrow (A \uparrow B)$$

$$A + B = (A \uparrow A) \uparrow (B \uparrow B)$$

I have some supplemental exercises using *Mathematica* to verify Jaynes's assertions about `NAND`.

Later, he shows how Bayes's Theorem implements elementary syllogisms from Classical Logic. This certainly was a revelation to me of how probability could, in fact, generalize logic. Building on Jaynes's lead, I take up this issue in earnest in Chapters Six and Seven.

2.10   Solved Exercises for Chapter Two

Exercise 2.10.1: Write out any syntactically correct Boolean formula using the new notation for Classical Logic.

Solution to Exercise 2.10.1

Here is a legitimate formula in the new notation illustrating the use of parentheses to indicate the order in which the operations should be carried out.

$$G(A, B) = \overline{B} \wedge ((A \vee B) \wedge (B \wedge \overline{A}))$$

The $\wedge$ operation on B and $\overline{A}$ takes place first, and the $\vee$ operation on A and B takes place next. Then, the $\wedge$ operates on the result of the first two operations. The $\wedge$ operation of $\overline{B}$ with the last result contained within the double parentheses is the final operation. The only possible result of any of these operations is T or F.

Exercise 2.10.2: Write out the analog to the above formula in the generic notation of Chapter One.

Solution to Exercise 2.10.2

The $\wedge$ operator is the analog to $\circ$, and the $\vee$ operator is the analog to $\bullet$. Variables A and B are the analogs to variables x and y. $\overline{A}$ and $\overline{B}$ are the analogs to x' and y'.

$$f(x, y) = y' \circ ((x \bullet y) \circ (y \circ x'))$$

Exercise 2.10.3: Express the same formula using the Classical Logic names for the operators. Use a prefix notation.

Solution to Exercise 2.10.3

Using *Mathematica* syntax (see Appendix A), we can write the above formula as,

And[Not[q], And[Or[p, q], And[q, Not[p]]]]

Exercise 2.10.4: Substitute some acceptable values for the variables in both the Classical Logic formula and the generic Boolean formula.

Solution to Exercise 2.10.4

Substitute the values of $A = T$ and $B = F$ for variables A and B in the Classical Logic formula.

$$G(A = T, B = F) = T \wedge ((T \vee F) \wedge (F \wedge F))$$

Substitute the values of $x = a'$ and $y = T$ for variables x and y in the above Boolean formula.

$$f(x = a', y = T) = F \circ ((a' \bullet T) \circ (T \circ a))$$

There are only two possible values, T or F, for any Classical Logic variable. But a Boolean formula may contain any number of values from its carrier set $\mathbf{B}$.

Exercise 2.10.5: Refer back to Tables 2.1 and 2.2, the binary operator tables for $\wedge$ and $\vee$, to solve for the particular settings of the variables in the Classical Logic formula $G(T, F)$ as given in the previous exercise.

Solution to Exercise 2.10.5

This formula has a value of F. Like any Boolean Algebra, Classical Logic is a "closed" system, so the value of any formula in Classical Logic must return a value of T or F.

$$T \vee F \quad \rightarrow \quad T \qquad \text{First inner parentheses}$$

$$F \wedge F \quad \rightarrow \quad F \qquad \text{Second inner parentheses}$$

$$T \wedge F \quad \rightarrow \quad F \qquad \text{Outer parentheses}$$

$$T \wedge F \quad \rightarrow \quad F \qquad \text{First term with outer parentheses}$$

Exercise 2.10.6: Show binary operator tables for the first two functions f_1 and f_2 from the sixteen possible Classical Logic functions.

Solution to Exercise 2.10.6

Table 2.11 reformats the definitions given in Tables 2.3 and 2.4 into binary operator tables.

Table 2.11: *The definition of the two binary operators $\perp$ and $\downarrow$ comparable to the similar tables for the $\wedge$ and $\vee$ operators.*

$\perp$	TRUE	FALSE
TRUE	FALSE	FALSE
FALSE	FALSE	FALSE

$\downarrow$	TRUE	FALSE
TRUE	FALSE	FALSE
FALSE	FALSE	TRUE

Exercise 2.10.7: Show the functional notation version for the Classical Logic formula $(A \perp B) \downarrow (A \downarrow B)$.

Solution to Exercise 2.10.7

Consult Table 2.5 to match up the operator symbols with the correct functions to write,

$$f_2 \,[\, f_1(A, B), f_2(A, B) \,]$$

Note that the outer function f_2 has two arguments, $f_1(A, B)$ and $f_2(A, B)$, each of which also has two arguments. Every function call must evaluate to T or F.

Exercise 2.10.8: Using the two binary operator tables just derived and shown in Table 2.11, solve the expression in Exercise 2.10.7 when the variables assume the value of $A = F$ and $B = F$.

Solution to Exercise 2.10.8

The solution is F.

$$
\begin{array}{rcll}
A \perp B \equiv f_1(F, F) & \rightarrow & F & \qquad \text{First term within brackets} \\[2mm]
A \downarrow B \equiv f_2(F, F) & \rightarrow & T & \qquad \text{Second term within brackets} \\[2mm]
f_2(F, T) & \rightarrow & F & \qquad \text{Outer function}
\end{array}
$$

Exercise 2.10.9: Translate the following Prefix formula using the names for the logic operators from Table 2.5, NAND [DIFFERENCE [A, B], EQUAL [A, B]], into the comparable infix formula and then into the functional notation.

Solution to Exercise 2.10.9

The infix notation using the symbols is,

$$(A \star B) \uparrow (A \leftrightarrow B)$$

while the functional notation is,

$$f_{12} \,[\, f_3(A, B), f_8(A, B) \,]$$

Exercise 2.10.10: If variable *A* is set at *F* and variable *B* is set at *T*, what is the solution to the last exercise?

Solution to Exercise 2.10.10

The composite function has a value of T.

$$f_3(F, T) \quad \rightarrow \quad T \qquad \text{First term within brackets}$$

$$f_8(F, T) \quad \rightarrow \quad F \qquad \text{Second term within brackets}$$

$$f_{12}(T, F) \quad \rightarrow \quad T \qquad \text{Outer function}$$

Exercise 2.10.11: Demonstrate to yourself that you can write at will the most complicated looking logic expression using the sixteen possible functions.

Solution to Exercise 2.10.11

Here is a little recipe with an example for constructing complicated looking logic expressions for two variables.

1. Pick any of the sixteen functional symbols. $\qquad\qquad\qquad\qquad \downarrow$

2. Place it between A and B. $\qquad\qquad\qquad\qquad A \downarrow B$

3. Wrap parentheses around this first expression. $\qquad\qquad (A \downarrow B)$

4. Pick another of the sixteen functional symbols. $\qquad\qquad \leftarrow$

5. Place it between A and B. $\qquad\qquad\qquad\qquad A \leftarrow B$

6. Wrap parentheses around this second expression. $\qquad (A \leftarrow B)$

7. Pick another of the sixteen functional symbols. $\qquad\qquad \lor$

8. Place it between the two expressions already constructed. $(A \downarrow B) \lor (A \leftarrow B)$

9. Make what has been constructed so far a new unit by wrapping it in another set of parentheses. $\qquad\qquad\qquad ((A \downarrow B) \lor (A \leftarrow B))$

10. Pick another one of the sixteen functional symbols. $\qquad\qquad \oplus$

11. Insert this new symbol between A and the new unit. $A \oplus ((A \downarrow B) \lor (A \leftarrow B))$

12. Continue for as long as you like.

Exercise 2.10.12: Prove that $A \triangleright B$ is logically equivalent to $(A \wedge B) \vee (\overline{A} \wedge B)$.

Solution to Exercise 2.10.12

The functional assignment for $A \triangleright B$ at each variable setting must match up with the functional assignments to,

$$(A \wedge B) \vee (\overline{A} \wedge B)$$

at all four possible variable assignments. The four possible variable assignments are, of course,

1. $A = T$ and $B = T$

2. $A = T$ and $B = F$

3. $A = F$ and $B = T$

4. $A = F$ and $B = F$

The functional assignments from **B** to $A \triangleright B$, as shown in Table 2.5 at each of these four settings, are, respectively, T, F, T, F. It is easy to check that when substituting these variable settings into the DNF expression $(A \wedge B) \vee (\overline{A} \wedge B)$ we get the same answer.

1. $A = T$ and $B = T$ $(T \wedge T) \vee (F \wedge T) \to T \vee F \to$ $\boxed{T}$

2. $A = T$ and $B = F$ $(T \wedge F) \vee (F \wedge F) \to F \vee F \to$ $\boxed{F}$

3. $A = F$ and $B = T$ $(F \wedge T) \vee (T \wedge T) \to F \vee T \to$ $\boxed{T}$

4. $A = F$ and $B = F$ $(F \wedge F) \vee (T \wedge F) \to F \vee F \to$ $\boxed{F}$

Now from the definition of a logical tautology we know that,

$$(A \triangleright B) \leftrightarrow ((A \wedge B) \vee (\overline{A} \wedge B))$$

must work out to T for all four settings. But from the definition of the `EQUAL` operator we know that if the first argument is the same as the second argument then the function returns a T. Refer back to Table 2.5 to verify this.

The left hand side of the $\leftrightarrow$ operator, the expression $(A \triangleright B)$ is $TFTF$, while the right hand side of the $\leftrightarrow$ operator, the expression $(A \wedge B) \vee (\overline{A} \wedge B)$ is also $TFTF$. Thus, the $\leftrightarrow$ operator will always return T. We have proven that the DNF expression as given above is logically equivalent to the $\triangleright$ operator.

Similarly, all sixteen binary operators should be logically equivalent to their DNF expressions. The important point is that any one of the 16 functions can be expressed using only the $\wedge$ and $\vee$ operators together with negation shown as the overbar.

Exercise 2.10.13: Show that function f_6 deserves its name.

Solution to Exercise 2.10.13

The name given to function f_6 was `NOT A`. To begin, prove a theorem in Boolean Algebra,

$$(x' \circ y) \bullet (x' \circ y') \rightarrow x'.$$

1. $(x' \circ y) \bullet (x' \circ y')$ Given

2. $x' \circ (y \bullet y')$ Distributivity (backwards)

3. $x' \circ T$ Complementation

4. x' Special Elements

Second, translate this theorem over to the notation for Classical Logic,

$$(\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B}) \rightarrow \overline{A}$$

Thus, the left hand side, which is the DNF representation for function f_6, reduces to the negation of A. Translating the overbar symbol into the infix symbols, the expression for the left hand side becomes,

$$((A \vdash B) \wedge B) \vee ((A \vdash B) \wedge (A \dashv B))$$

Exercise 2.10.14: Simplify the full DNF expression for function f_{15}.

Solution to Exercise 2.10.14

Function f_{15} is the `OR` operator. Its DNF representation from Table 2.6 is

$$(A \wedge B) \vee (A \wedge \overline{B}) \vee (\overline{A} \wedge B)$$

1. $(A \wedge B) \vee (A \wedge \overline{B}) \vee (\overline{A} \wedge B)$ Given

2. $(A \wedge B) \vee (\overline{A} \wedge B) \vee (A \wedge \overline{B})$ Reorder second and third terms

3. $(B \wedge (A \vee \overline{A})) \vee (A \wedge \overline{B})$ Distributivity (backwards) and Commutativity on first two terms

4. $(B \wedge T) \vee (A \wedge \overline{B})$ Complementation

5. $B \vee (A \wedge \overline{B})$ Special Elements

6. $(B \vee A) \wedge (B \vee \overline{B})$ Distributivity

7. $(A \vee B) \wedge T$ Commutativity and Complementation

8. $A \vee B$ Special Elements

In the end, we see that the DNF expression does, in fact, simplify to the `OR` operation between A and B.

Exercise 2.10.15: Show the Boolean functional assignment table for
$f_{11}(A, B, C)$.

Solution to Exercise 2.10.15

Following the pattern established by constructing $f_{10}(A, B, C)$ as explained earlier, the T functional assignment for f_{11} is moved up to the sixth position while the seventh position reverts back to an F as shown below in Table 2.12.

Table 2.12: *A Boolean functional assignment table for three variables illustrating the eleventh of the 256 possible logic functions.*

A	B	C	$f_{11}(A, B, C)$
T	T	T	F
T	T	F	F
T	F	T	F
T	F	F	F
F	T	T	F
F	T	F	T
F	F	T	F
F	F	F	T

Exercise 2.10.16: Construct the DNF expression for $f_{11}(A, B, C)$.

Solution to Exercise 2.10.16

Using the heuristic described in this Chapter, f_{11} takes on the value T for two terms. Thus, there will be two terms in the DNF expression with each one of these two terms consisting of a conjunction of the three variables and the two terms joined by a disjunction.

The first T is the sixth row where the variable setting is $A = F$, $B = T$, and $C = F$. Therefore, the first term is $\overline{A} \wedge B \wedge \overline{C}$. The second T occurs in the last row where the variable setting is $A = F$, $B = F$, and $C = F$. The second term, then, is $\overline{A} \wedge \overline{B} \wedge \overline{C}$. Joining these two conjunctive terms with the disjunctive operator yields the DNF expression for $f_{11}(A, B, C)$ as $\overline{A}B\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}$.

Exercise 2.10.17: Is $((A \to B) \oplus C) \leftrightarrow (A \vee B) \wedge \overline{C}$ a tautology?

Solution to Exercise 2.10.17

If we can show that this composite logic function evaluates to F for any one of the eight possible variable settings, then it is not a tautology. In other words,

$(A \rightarrow B) \oplus C$ is not logically equivalent to $(A \vee B) \wedge \overline{C}$. Let $A = T$, $B = F$, and $C = T$ be the third possible variable setting in the usual way that we present the variables. Then set up the functional version for the composite function with the appropriate arguments for this variable setting as,

$$f_8 \{ f_9 [f_{13}(T, F), T], f_5 \{ f_7 [f_{15}(T, F), T] \} \}$$

Now, working our way through each function from the inside out, we have,

$$
\begin{aligned}
f_{15}(T, F) &= T \\
f_7(T, T) &= F \\
f_5(F, T) &= F \\
f_{13}(T, F) &= F \\
f_9(F, T) &= T \\
f_8(T, F) &= F
\end{aligned}
$$

Thus, for this particular variable setting, the evaluation yields F. The logic expression in question,

$$((A \rightarrow B) \oplus C) \leftrightarrow (A \vee B) \wedge \overline{C}$$

is not therefore a tautology. There are two other variable settings for which the logic expression also yields an F. They are TFF and FFF. As you might have expected, I discovered these two other failures of the tautology with a *Mathematica* program, but the tedious working through by hand is satisfying in its own peculiar way.

Exercise 2.10.18: Solve the third "weird–looking" logic formula in section 2.4 for some given setting of the variables.

Solution to Exercise 2.10.18

Suppose that the setting for the variables was given as $A = F$ and $B = T$. After reviewing Table 2.5, match up the functional symbols with the correct $f_j(A, B)$ and substitute the values for A and B. Then, consult the correct column and read off the functional assignment. Start from the innermost set of parentheses on the right. See Appendix A for the short *Mathematica* program that solves this problem.

$$A \dashv B \;=\; f_7(F, T)$$

$$f_7(F, T) \;=\; F$$

$$A \vdash B \;=\; f_6(F, T)$$

$$f_6(F, T) \;=\; T$$

$$(A \vdash B) \vartriangleleft (A \dashv B) \;=\; f_{11}\,[\,f_6(F, T), f_7(F, T)\,]$$

$$f_{11}(T, F) \;=\; T$$

$$A \leftarrow B \;=\; f_{14}(F, T)$$

$$f_{14}(F, T) \;=\; F$$

$$A \diamond B \;=\; f_4(F, T)$$

$$f_4(F, T) \;=\; F$$

$$(A \diamond B) \vartriangleright (A \leftarrow B) \;=\; f_{10}\,[\,f_4(F, T), f_{14}(F, T)\,]$$

$$f_{10}(F, F) \;=\; F$$

$$((A \diamond B) \vartriangleright (A \leftarrow B)) \downarrow ((A \vdash B) \vartriangleleft (A \dashv B)) \;=\; f_2(F, T)$$

$$f_2(F, T) \;=\; F$$

The notational translation as we move through operator syntax to functional syntax, and finally to *Mathematica* syntax is,

1. $((A \diamond B) \vartriangleright (A \leftarrow B)) \downarrow ((A \vdash B) \vartriangleleft (A \dashv B))$

2. $f_2\,\{f_{10}\,[\,f_4(A, B), f_{14}(A, B)\,], f_{11}\,[\,f_6(A, B), f_7(A, B)\,]\,\}$

3. **`Nor[f10[f4[p, q], f14[p, q]], f11[f6[p, q], f7[p, q]]]`**

Chapter 3

Cellular Automata

3.1 Introduction

We devote an entire Chapter to a specific application using a three variable Boolean function, or more specifically, a logic function with three arguments. The example studied here introduces the so–called *elementary cellular automaton.* These Cellular Automata[1] are very widely used in computational theory, and one of our objectives is to see if they can be generalized by probability theory.

Wolfram does an excellent job of stressing just how important cellular automata can be in our understanding of many phenomena. To this end, he presents an astounding example of one very simple cellular automaton which he calls "Rule 110."

The reason why Rule 110 is so remarkable is because it emulates a Universal Turing Machine. Due to Rule 110's universality, this CA can in principle compute anything that any other computational device can compute, no matter how complex that "anything" might be. Because Wolfram uses a particular Boolean function of three variables for his Rule 110, we will examine this function in some detail.

Rule 110 is still a *deductive* process because there is no uncertainty as to the functional assignment for the three variables, Later, we will want to generalize this deductive process into an *inferential* process. In other words, we would like to admit to some uncertainty in the output of the CA due to the lack of some relevant information.

So, just like Boolean Algebra and Classical Logic, Cellular Automata are strictly deductive phenomena. We would never need to invoke information processing, inference, and probability if we were to limit our curiosity to these pursuits.

---

[1] In the text, both Cellular Automaton and Cellular Automata are often abbreviated to CA.

3.2   Rule 110 as a Boolean Function

The general abstract definition for a Boolean function of three variables helps to understand this cellular automaton. The template for the construction of a three variable function was studied at the end of the last Chapter and given as,

$$f : \mathbf{B}^3 \rightarrow \mathbf{B}$$

The set $\mathbf{B}$ is defined as before, consisting of only the special elements T and F. There are eight elements in the domain. Thus the set,

$$\mathbf{B} \times \mathbf{B} \times \mathbf{B} = \{\, TTT, TTF, TFT, TFF, FTT, FTF, FFT, FFF \,\}$$

is composed of ordered triples. Each one of these ordered triples will be assigned some element from $\mathbf{B}$ by the function f. So, for example, we might write some function, with its arguments being the first element in the domain, as $f_\star(TTT) = F$.

For the $n = 3$ variable Boolean functions, there are a total of $2^{2^3} = 256$ possible functional assignments where an F or a T is assigned to all eight possible values that the three variables can assume. Table 3.1 presents *one* of these 256 possible functions showing the assignment to each possible setting of the three variables. In fact, this is the functional assignment which defines Rule 110.

Table 3.1: *The functional assignment which is Rule 110.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
F	T	T	F	T	T	T	F

It is possible to write this function, like any other logic function, in its disjunctive normal form (DNF). Wherever the function assumes the value T, we look to see which of the three variables is T and which are F.

So for the first term in the DNF for Rule 110, we find the first T in the bottom row of Table 3.1 and write down $A \wedge B \wedge \overline{C} \equiv AB\overline{C}$ to match up with TTF appearing above that first T.

There are a total of five Ts in the bottom row of Table 3.1 and doing the same thing for each one of these five Ts, we find that the DNF for Rule 110 consists of five terms each connected by a disjunction, that is, the $\vee$ (OR) binary operator. Each one of the five terms is itself composed entirely of the $\wedge$ operator, shown below in condensed notation.

$$f_{110}(A, B, C) = AB\overline{C} \vee A\overline{B}C \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$$

3.3 How a Cellular Automaton Evolves

A cellular automaton shows how some ontological system defined by a rule evolves over time. The simplest cellular automata are called *one-dimensional* because they begin with a single line of cells. This starting line of cells is allowed to be arbitrarily long. These cells are colored either black or white.

The beginning line of cells is transformed at each time step into another line of cells where, again, each cell is black or white. The transformation of each cell takes place at distinct time steps by following some rule which looks at the cell's neighbors at the preceding time step. See Figure 3.1 for a schematic drawing of a CA as it evolves over time by following Rule 110.
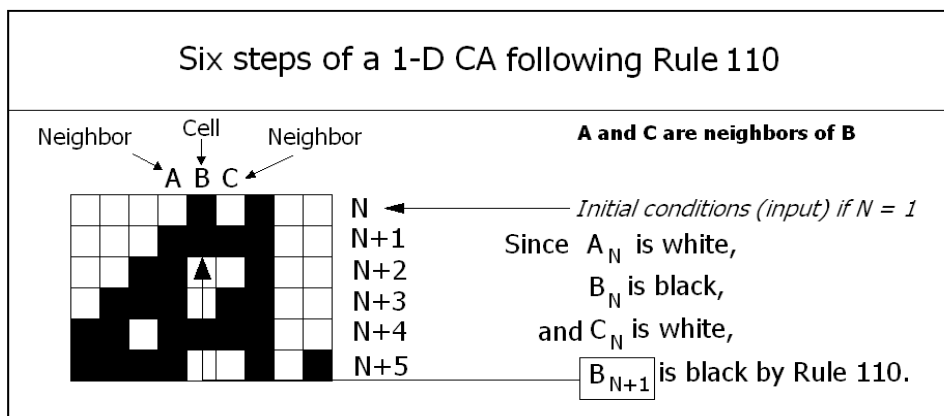
Figure 3.1: *The evolution of a cellular automaton following Rule 110 over five time steps.*

Suppose that we begin by finding the cellular automaton at some arbitrary time step N with some number of cells colored either white or black. The CA evolves over time by applying a particular rule to the colors of certain cells at the previous time step. Rule 110 looks at the color of just three cells in the immediately previous time step.

This is the cell above the current cell as well as that cell's right and left neighbors. These three cells are labeled as A_N, B_N, and C_N with cell B_N immediately above the cell B_{N+1} scheduled for the update. Cell A_N is the left neighbor of B_N and cell C_N is the right neighbor of B_N. Depending on the color of these three cells, cell B_{N+1} will be colored either black or white at the next time step.

For example, at time $N + 1$, cell B_{N+1} looks at the cells A_N, B_N, and C_N, that is, the cell immediately above it and that cell's two neighbors. If A_N happens to be white, B_N black, and C_N white, then B_{N+1} is updated to black by Rule 110. One can imagine applying such a rule starting at the left end cell and then marching through each cell in turn until the cell at the right end is reached. There has to be

some convention on how to treat the cells at either end because one of the neighbors will be missing.

In fact, this Rule 110 algorithm for coloring each current time step's cell is none other than the Boolean function shown in Table 3.1. Let T stand for *black* and F for *white*. The possible variable settings in the first row of this table are the colors of the A_N, B_N, and C_N cells at the immediately preceding time step. The functional assignment in the second row is then the color for B_{N+1}.

So, for example, the first column in Table 3.1 tells us that if the above three cells are all black, then the cell in question is white. In the initial explanatory example, the cells above B_{N+1} were white, black, and white, and therefore B_{N+1} was colored black. The sixth column of the functional assignment table implements this because $f_{110}(F, T, F) = T$. See Figure 3.2 which duplicates the previous CA, but now explicitly references Rule 110.
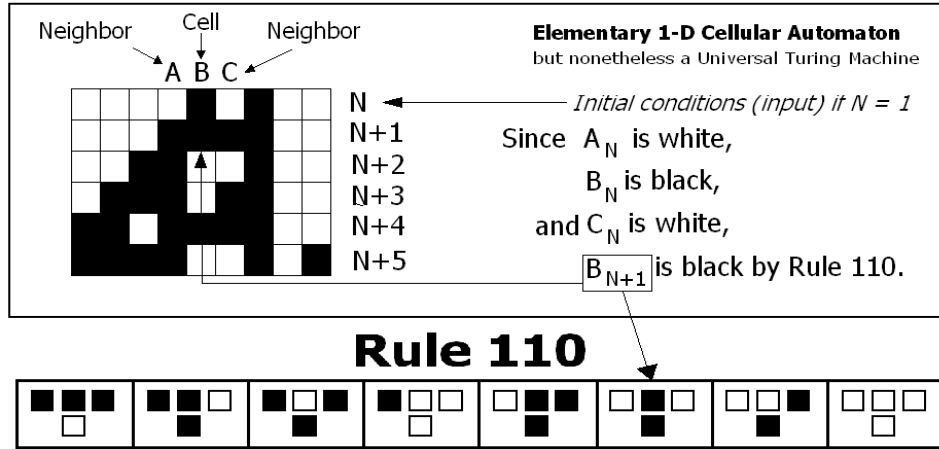
Figure 3.2: *The evolution of a cellular automaton following Rule 110 over five time steps. The specific case in Rule 110 dictating B_{N+1}'s updated color is shown.*

3.3.1   Wolfram's numbering scheme for rules

If one were to express the functional assignment shown in Table 3.1 (second row) in base 2 where $F = 0$ and $T = 1$, then we have,

$$
\begin{aligned}
01101110 &= (0 \times 128) + (1 \times 64) + (1 \times 32) + (0 \times 16) + \\
&\quad (1 \times 8) + (1 \times 4) + (1 \times 2) + (0 \times 1) \\
&= 110
\end{aligned}
$$

Later on, we will deal with an example that has the functional assignment as shown in Table 3.2.

Table 3.2: *The functional assignment which is Rule 192.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
T	T	F	F	F	F	F	F

$$
\begin{aligned}
11000000 &= (1 \times 128) + (1 \times 64) + (0 \times 32) + (0 \times 16) + \\
&\quad (0 \times 8) + (0 \times 4) + (0 \times 2) + (0 \times 1) \\
&= 192
\end{aligned}
$$

Thus, this functional assignment for a three variable Boolean function which might be used to update a CA is called Rule 192.

3.4   Rule 110 in Different Forms

When we express a Boolean function in its full disjunctive normal form, it does not mean that the expression could not be written in different, or simpler forms. Remember, there are an infinite variety of Boolean formulas for any given Boolean function. We will illustrate this with Rule 110.

Here, once again, is the DNF expansion for Rule 110 with its five terms,

$$
\overbrace{AB\overline{C}}^{1} \vee \overbrace{A\overline{B}C}^{2} \vee \overbrace{\overline{A}BC}^{3} \vee \overbrace{\overline{A}B\overline{C}}^{4} \vee \overbrace{\overline{A}\,\overline{B}C}^{5}
$$

where, for the ensuing discussion, we have explicitly labeled each of the five terms. Now, reorder these five terms as follows,

$$
\overbrace{AB\overline{C}}^{1} \vee \overbrace{\overline{A}B\overline{C}}^{4} \vee \overbrace{A\overline{B}C}^{2} \vee \overbrace{\overline{A}\,\overline{B}C}^{5} \vee \overbrace{\overline{A}BC}^{3}
$$

because we want to set things up for factoring. Here is the tricky part. Because of the **Idempotence axiom**, any logical expression is not changed by ORing with a term already in the expression. In other words, $x \bullet x = x$. Therefore, add to the already existing five terms another sixth term, $\overline{A}B\overline{C}$, which repeats the fourth term.

$$
\overbrace{AB\overline{C}}^{1} \vee \overbrace{\overline{A}B\overline{C}}^{4} \vee \overbrace{A\overline{B}C}^{2} \vee \overbrace{\overline{A}\,\overline{B}C}^{5} \vee \overbrace{\overline{A}BC}^{3} \vee \overbrace{\overline{A}B\overline{C}}^{6=4}
$$

Factor out what is common when taking two terms at a time in order,

$$
\begin{aligned}
AB\overline{C} \vee \overline{A}B\overline{C} &= B\overline{C}\,(A \vee \overline{A}) \\[2mm]
&= B\overline{C} \\[2mm]
A\overline{B}C \vee \overline{A}\,\overline{B}C &= \overline{B}C\,(A \vee \overline{A}) \\[2mm]
&= \overline{B}C \\[2mm]
\overline{A}BC \vee \overline{A}B\overline{C} &= \overline{A}B\,(C \vee \overline{C}) \\[2mm]
&= \overline{A}B
\end{aligned}
$$

which means that the original DNF can now be re–written as,

$$
B\overline{C} \vee \overline{B}C \vee \overline{A}B
$$

Wolfram provides several other logic expressions for Rule 110. However, they are not written in the form of the simplified DNF expression as just derived either. For example, Wolfram gives the following logic expression (in our notation) for Rule 110,

$$
(\overline{A} \wedge B \wedge C) \oplus B \oplus C
$$

However, the two expressions are logically equivalent. A small program shows that

$$
(\overline{A}BC \oplus B \oplus C) \leftrightarrow (B\overline{C} \vee \overline{B}C \vee \overline{A}B)
$$

evaluates to T for all eight possible variable settings, therefore, the two expressions are the same (logically equivalent). One of the eight variable settings is solved in detail at the end of the Chapter using the functional notation of Chapter Two.

3.5   Another Example of a Cellular Automaton

It was mentioned that there are 256 possible functions for three variables where the carrier set **B** consists solely of the elements T and F. Rule 110 was interpreted as one of these 256 functions. A one–dimensional cellular automaton consisting of two colors and updating according to the colors of three cells at the previous time step could follow any one of these 256 functions.

Figure 3.1 showed a small segment of a CA following Rule 110. Here is another example in Figure 3.3, appearing at the top of the next page, of a CA that follows Rule 192.

Earlier, we showed the translation from Wolfram's numbering scheme to the Boolean functional assignment. The DNF for Rule 192 is particularly easy to write down.
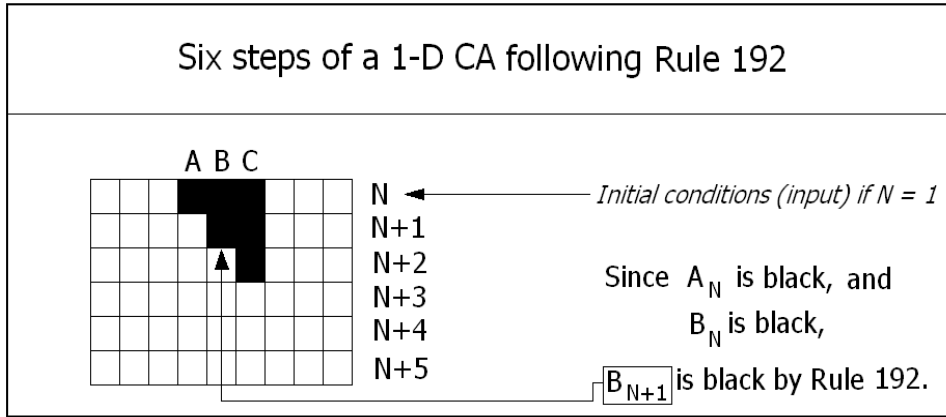
Figure 3.3: *The evolution of a cellular automaton following Rule 192 over five time steps.*

It consists of the two terms where T is located in the second row of Table 3.2,

$$ABC \vee AB\overline{C}$$

This full DNF expression is, in turn, easily reduced to,

$$ABC \vee AB\overline{C} \quad = \quad AB \, (C \vee \overline{C})$$

$$= \quad AB$$

So, the cell at the next time step, B_{N+1}, will be colored black only if the cell above it, B_N, and its left neighbor, A_N, are both black.

One of the important properties of a CA following Rule 110 is that it generates "localized structures" which then interact with other structures. These localized structures also persist for some finite time as the CA evolves. It was only through the complicated interaction of these localized structures that Wolfram could prove that Rule 110 was a Universal Turing Machine.

However, in a CA following Rule 192 all the black cells will eventually disappear along with any localized structures. Therefore, a CA following Rule 192 exhibits perfectly predictable behavior. For Wolfram, the inherent *unpredictability* of Rule 110 places it into a much more interesting category than the inherent *predictability* of Rule 192.

For us, what is predictable or not predictable is viewed through the lens of probability theory. We will present what we believe is a more nuanced look at predictability when we recast cellular automata within a probabilistic framework later on in the book.

3.6   Generalizations of These Elementary CA

Now it should be clear that these notions about cellular automata can be greatly generalized. One step in generalizing would be to allow n-variable Boolean functions to act as rules. So, for example, consider CA which can update by looking at *two* left and right neighbors above the cell to be updated, instead of just one neighbor. The notion of only two colors allowed for any cell is still retained.

Another direction in which to generalize would be to relax this requirement for only two colors. In addition to allowing cells to update to black or white, a cell might be allowed to update to two additional colors, say, dark gray and light gray.

3.6.1   A five variable function

Let's extend the notion of a three variable function as a rule assignment in a CA to a five variable function. First, we fill in the template for a Boolean function of five variables, $f : \mathbf{B}^5 \to \mathbf{B}$.

There is an ordered quintuple of five elements from $\mathbf{B}^5$ to which is assigned a T or an F. For example, $f_\star(TFTFT) = F$. The T is interpreted as a black cell and the F as a white cell, just like the last time. The number of possible rules jumps up enormously from $2^{2^3} = 256$ rules to over four billion rules, $2^{2^5} = 4,294,967,296$.

To keep things manageable, consider Rule 2,147,483,649. This rule has a 1 at positions $2^{31} = 2,147,483,648$ and $2^0 = 1$ with 0s everywhere else. Thus, a T occurs only for the two variable settings of $TTTTT$ and $FFFFF$ from the possible 32 settings. The DNF is $ABCDE \vee \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}\,\overline{E}$.

Let's label the cell to be updated as C_{N+1}, with the two left neighbors at the previous time step being A_N and B_N and the two right neighbors being D_N and E_N. So, for a CA following Rule 2,147,483,649, cell C_{N+1} gets updated to *black* only if the cell above and all four neighbors are black, or if the cell above and all four of its neighbors are white. In all other cases, the cell gets updated to *white*.

3.6.2   More than two colors

Let's combine the generalization of the last section with the generalization of adding two more colors. Thus, we are thinking about a CA which updates the current cell by looking at the cell above and its *two* neighbors on each side. These five cells might be colored white, black, light gray, or dark gray. If we let k stand for the number of colors and n for the number of variables, then the formula for the number of rules is k^{k^n}, which, for the current example, is a staggering $4^{4^5} = 4^{1024} \approx 3.23 \times 10^{616}$.

This serves as a good example of what is called a *combinatorial explosion*. The enormous numbers involved stand as a major impediment to any kind of direct analysis. We shall experience the unpleasant nature of combinatorial explosions

later on in the book. The wide spread prevalence of combinatorial explosions is another inducement for inferential reasoning as opposed to pure deduction.

Since we have expanded to four colors, let the carrier set consist of elements $\mathbf{B} = \{a, a', F, T\}$ where a is associated with light gray, a' with dark gray, and, as before, F with white and T with black. Since we are dealing with a carrier set consisting of more than F and T, let's revert back to the original variable notation of Chapter One.

Suppose, somewhat arbitrarily, that the rule in question for this CA is the Boolean formula, $f(v, w, x, y, z) = (avw) \bullet (a'x'y'z')$ where the usual convention is employed of hiding the $\circ$ operation.

Suppose further, in order to advance this little example, that the color of the five cells at time step N relevant to updating a cell at time step $N+1$ are, in order, light gray, black, white, dark gray, and black. Thus, matching up light gray with a, black with T, white with F, dark gray with a', and black again with T, the arguments to the function result in,

$$
\begin{aligned}
f(a, T, F, a', T) &= ((a \circ a) \circ T) \bullet (((a' \circ T) \circ a) \circ F) \\[2mm]
&= (a \circ T) \bullet ((a' \circ a) \circ F) \\[2mm]
&= a \bullet (F \circ F) \\[2mm]
&= a \bullet F \\[2mm]
&= a
\end{aligned}
$$

Therefore, cell C_{N+1} will be colored light gray according to this rule.

3.7   Solved Exercises for Chapter Three

Exercise 3.7.1: Use the top row of Table 3.1 as a template and fill in the second row arbitrarily with *T*s and *F*s.

Solution to Exercise 3.7.1

I arbitrarily filled in the second row by alternating *T*s and *F*s.

Table 3.3: *An arbitrary functional assignment which is some CA rule.*

TTT	*TTF*	*TFT*	*TFF*	*FTT*	*FTF*	*FFT*	*FFF*
T	*F*	*T*	*F*	*T*	*F*	*T*	*F*

Exercise 3.7.2: Is whatever pattern you just filled in some function?

Solution to Exercise 3.7.2

Yes, it is. It is a rule for associating each ordered triple from $\mathbf{B} \times \mathbf{B} \times \mathbf{B}$ with an element from $\mathbf{B}$.

Exercise 3.7.3: What is the rule number in Wolfram's numbering scheme for the logic function you just created?

Solution to Exercise 3.7.3

Match up each T with a 1 and each F with a 0 in the typical decomposition for a binary number starting with 2^7 and ending with 2^0.

$$(128 \times 1) + (64 \times 0) + (32 \times 1) + (16 \times 0) + (8 \times 1) + (4 \times 0) + (2 \times 1) + (1 \times 0) = 170$$

Therefore, I just created Rule 170.

Exercise 3.7.4: Use Wolfram's numbering scheme to decode Rule 124.

Solution to Exercise 3.7.4

In the binary system,

$$124 = (128 \times 0) + (64 \times 1) + (32 \times 1) + (16 \times 1) + (8 \times 1) + (4 \times 1) + (2 \times 0) + (1 \times 0)$$

Therefore, the pattern of *T*s and *F*s corresponding to the 1s and 0s are matched up with all eight variable settings to yield,

Table 3.4: *The functional assignment which is Rule 124.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
F	T	T	T	T	T	F	F

Exercise 3.7.5: Using the results of the previous exercise, what is the DNF expansion of Rule 124?

Solution to Exercise 3.7.5

Looking at where the five Ts occur in the bottom row, the DNF expansion for Rule 124 consists of the five terms,

$$AB\overline{C} \vee A\overline{B}C \vee A\overline{B}\,\overline{C} \vee \overline{A}BC \vee \overline{A}B\overline{C}$$

Exercise 3.7.6: Is the logic expression $(A \oplus (A \wedge B \wedge \overline{C})) \oplus B$ a formula for Rule 124?

Solution to Exercise 3.7.6

Yes, it is. The functional assignments, in order, to all eight variable settings for this logical expression are $FTTTTTFF$, the same as Rule 124. A *Mathematica* program in Appendix B will prove this for all variables settings, but, for now, hand calculate the particular variable setting of $A = F$, $B = F$, and $C = T$. This is FFT, the seventh, or next to last variable setting in the standard order in which we present them. This variable setting should result in a functional assignment of F. In other words, $f_{124}(F, F, T) = F$. In the rather hard to follow functional notation we set up in Chapter Two, evaluate the expression as,

$$
\begin{aligned}
(A \oplus (A \wedge B \wedge \overline{C})) \oplus B &= f_9\,[\,f_9\,\{\,F, f_5\,[\,f_5(F,F), F\,]\,\}, F\,] \\[2mm]
&= f_9\,[\,f_9\,\{F, f_5(F,F)\,\}, F\,] \\[2mm]
&= f_9\,[\,f_9(F,F), F\,] \\[2mm]
&= f_9(F,F) \\[2mm]
&= F
\end{aligned}
$$

Exercise 3.7.7: Is the logic expression $(A \vee B) \oplus (A \wedge B \wedge C)$ another formula for Rule 124?

Solution to Exercise 3.7.7

Yes, it is. Once again, the functional assignments to all eight variable settings for this logic expression are $FTTTTTFF$, and match up with Rule 124. The hand calculation for a variable setting of $A = T$, $B = T$, and $C = F$ should result in a functional assignment of T. In other words, $f_{124}(T, T, F) = T$. In the easier to follow *Mathematica* syntax, $(A \vee B) \oplus (A \wedge B \wedge C)$ gets translated into,

Xor[Or[True, True], And[True, True, False]]

which *Mathematica* then evaluates as **True**.

Exercise 3.7.8: Show the next step, step $N + 6$, in the evolution of the Rule 110 cellular automaton with a different starting configuration.

Solution to Exercise 3.7.8

The starting configuration of the cell colors at time N is white, white, white, white, black, white, black, white, and white. Thus, we show only a finite number of cells starting off the CA. By convention, assume that there is a "virtual" black cell adjoining the beginning cell at the left border and the end cell at the right border at step $N + 5$. This is because the convention is to "wrap around" the color at the end cell as the virtual color before the beginning cell and vice versa. Then simply follow the rules for filling in the color of a cell at step $N + 6$ by looking at the colors of the cell above and its two neighbors.

We make things a bit easier by constructing the analog to Table 3.1, Table 3.5 as shown below, with colors substituted for T and F. So, starting at the left most cell for step $N + 6$, BBB produces W, and so forth, until the final right cell is colored B from WBB. The current state of the Rule 110 CA at step $N + 6$ looks like the sketch shown in Figure 3.4.

Table 3.5: *Rule 110 with colors substituted for T and F.*

BBB	BBW	BWB	BWW	WBB	WBW	WWB	WWW
W	B	B	W	B	B	B	W
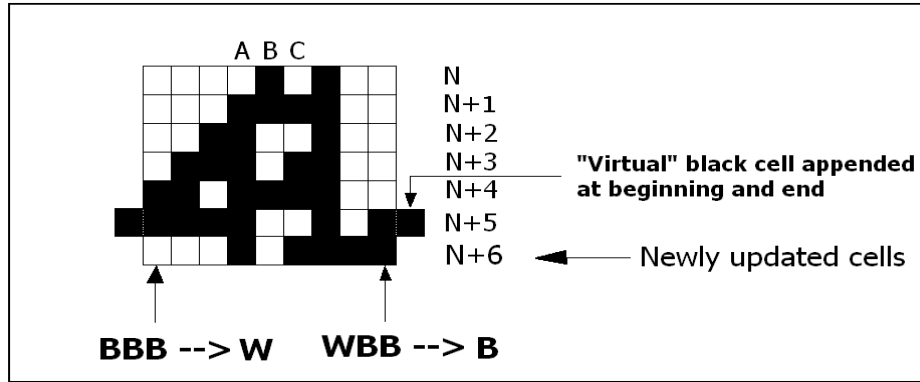
Figure 3.4: *The next time step evolution of a CA following Rule 110.*

Exercise 3.7.9: Show that one of Wolfram's alternative logic expressions for Rule 110 is logically equivalent to the simplified DNF for Rule 110.

Solution to Exercise 3.7.9

It was mentioned previously that Wolfram provides this logic expression, among others, for Rule 110,

$$(\overline{A} \wedge B \wedge C) \oplus B \oplus C$$

and the claim was made that this expression was the same as the simplification carried out for the full DNF expression. Thus, we have to prove logical equivalency between the two expressions,

$$((\overline{A} \wedge B \wedge C) \oplus B \oplus C) \leftrightarrow (B \wedge \overline{C}) \vee (\overline{B} \wedge C) \vee (\overline{A} \wedge B)$$

To prove logical equivalency, the left hand side and the right hand side of the EQUAL operator must be the same for all *eight* possible combinations of the variables. In this exercise we will illustrate the solution for one case where $A = T$, $B = F$, and $C = T$. The full solution is provided in Appendix B.

First, we evaluate the left hand side and then the right hand side. The left hand side, Wolfram's expression, evaluates to T given the settings for the variables.

$$f_9(f_9(f_5(f_5(F,F),T),F),T) = f_9(f_9(f_5(F,T),F),T)$$

$$f_9(f_9(f_5(F,T),F),T) = f_9(f_9(F,F),T)$$

$$f_9(f_9(F,F),T) = f_9(F,T)$$

$$f_9(F,T) = T$$

The right hand side, the DNF simplification, should also evaluate to T if the $\leftrightarrow$ operator operating on the results of left hand side and the right hand side is going

to report back a T.

$$f_{15}(f_{15}(f_5(F,F),f_5(T,T)),f_5(F,F)) \quad = \quad f_{15}(f_{15}(F,T),f_5(F,F))$$

$$f_{15}(f_{15}(F,T),f_5(F,F)) \quad = \quad f_{15}(T,F)$$

$$f_{15}(T,F) \quad = \quad T$$

Both the right hand expression and the left hand expression evaluate to T and $f_8(T,T) = T$, so one of the eight possible variable combinations checks out.

Exercise 3.7.10: Sketch the evolution for the CA that we have called Rule 2,147,483,649 for a few time steps. The CA starts out under the initial conditions that the first five cells are black and the next five cells are white.

Solution to Exercise 3.7.10

This CA is a generalization of a CA following a three variable Boolean function. The rules now are equivalent to a five variable Boolean function. See Figure 3.5 for a sketch of the CA following Rule 2,147,483,649 for six time steps.

Figure 3.5: *The initial evolution of a CA following Rule 2,147,483,649. This rule is a five variable Boolean function which looks at the cell above as well as its two left and two right neighbors to determine the color of the cell to be updated.*

Only when the cell above has all four neighbors of the same color will the cell to be updated be colored black. In all other cases, the cell to be updated is colored white. The two "virtual" cells on the left border are white and the two "virtual" cells on the right border are black.

At the first time step, the CA evolves according to the rules and reduces the number of black cells. Then, at the second time step, all cells become white. But

now at the third time step, since all the cells are white, the CA undergoes a violent reversal and all of the cells are transformed back into black cells. Having reached a configuration of all black cells, the CA stops evolving. All subsequent times steps can do nothing except produce further black cells.

Exercise 3.7.11: During the discussion of a CA characterized by more than two colors, it was shown that the color of a cell was updated to light gray. What is the color of the cell to its right?

Solution to Exercise 3.7.11

The CA is following a five variable Boolean function from a carrier set,

$$\mathbf{B} = \{a, a', F, T\}$$

where, as in the simpler CA, F is associated with the color white and T with the color black. A cell can now have two other colors, light gray and dark gray, associated with a and a'.

The function, that is, the rule the CA follows to update a cell, was given as,

$$f(v, w, x, y, z) = (avw) \bullet (a'x'y'z')$$

At the end of the Chapter, an example was given where a cell was updated to light gray.

If we move the updating process one cell to the right, everything is shifted one cell to the right. The cell above is now dark gray, the two neighbors to the left are black and white, while the two neighbors to the right are black and light gray. The furthest neighbor on the right, cell E, is a new cell and suppose it is, in fact, light gray. Substituting these values into the formula yields,

$$
\begin{aligned}
f(T, F, a', T, a) &= (aTF) \bullet (a'aFa') \\[2mm]
&= F \bullet F \\[2mm]
&= F
\end{aligned}
$$

Therefore, this cell, C_{N+1}, will be updated to white. For a visual mooring, see Figure 3.6 at the top of the next page.

Exercise 3.7.12: (After reading Appendix B.) Using the same technique as in section 3.4, show that the DNF for Rule 110 can be simplified in another way using four terms instead of three.

Solution to Exercise 3.7.12

The original DNF can also be re–written with a fourth term as,

$$B\overline{C} \vee \overline{B}C \vee \overline{A}B \vee \overline{A}C$$
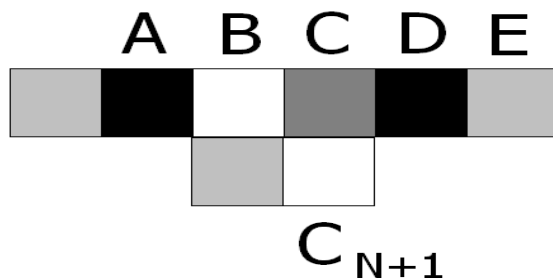
Figure 3.6: *Updating a cell's color in a cellular automaton with a generalization involving more neighbors and colors.*

In fact, this is the expression returned by,

FullForm[LogicalExpand[w1Rule110[p, q, r]]]

instead of the three term answer we derived earlier, even though the three term answer is returned for other ways of writing Rule 110. This longer expression is not wrong as we now demonstrate using the same tactics employed in section 3.4.

In section 3.4, a sixth term was added to the already existing five terms of the DNF. This addition repeated the fourth term. Then, factoring out the six term expression resulted in,

$$B\overline{C} \vee \overline{B}C \vee \overline{A}B$$

If two more terms already in the DNF are also added, then a total of eight terms will have been created. Factoring in the same way now results in a four term simplification of the original DNF. Terms 4, 3, and 5 are repeated,

$$\overbrace{AB\overline{C}}^{1} \vee \overbrace{\overline{A}B\overline{C}}^{4} \vee \overbrace{A\overline{B}C}^{2} \vee \overbrace{\overline{A}\,\overline{B}C}^{5} \vee \overbrace{\overline{A}BC}^{3} \vee \overbrace{\overline{A}B\overline{C}}^{4} \vee \overbrace{\overline{A}BC}^{3} \vee \overbrace{\overline{A}\,\overline{B}C}^{5}$$

Factor out a new term from the last two terms added,

$$\overline{A}BC \vee \overline{A}\,\overline{B}C \quad = \quad \overline{A}C\,(B \vee \overline{B})$$

$$= \quad \overline{A}C$$

so that the original DNF can now be re–written as,

$$B\overline{C} \vee \overline{B}C \vee \overline{A}B \vee \overline{A}C$$

This exercise is another example of the statement that there may be many different Boolean *formulas* for the same Boolean *function*.

Chapter 4

Analogies Between Formal Manipulations

4.1 Introduction

This Chapter serves as a prelude to the formal aspects of manipulating probability symbols. We want to show by small degrees how the formal manipulations of Boolean Algebra get translated over into analogous manipulations for probability symbols.

For a fruitful start on how Classical Logic can be extended by probabilities, we revert back to the generality inherent in a Boolean Algebra. In Chapter Two, during the discussion of Classical Logic, we restricted the carrier set $\mathbf{B}$ to just the two elements F and T. Thus, variables could only assume the values of T or F. But this also meant that the functional assignments had to be a T or F as well.

The assignment strategy must be broadened so that values other than T or F are available. This flexibility is a necessary requirement for understanding probability as an IP's uncertainty due to missing information.

If we examine a carrier set from a Boolean Algebra with more than two values, we begin the process that eventually leads to probability functions. It is important to note that we are still not emphasizing the aspects of how numerical values are attached to statements. We would like to tarry a while longer in the abstract mood with the expectation that such a dalliance will eventually make probability functions seem less alien.

However, it is the worst form of pedantic spite not to make things clearer when possible. To avoid unnecessary obfuscation, *joint probability tables* are employed extensively throughout my work starting with this Chapter. Thus, we start talking about probability before we have even defined it.

It is my intent to try a different approach than the one usually employed in introducing probability theory. I'd like to begin by merging gently from operations on one formal system to analogous operations in another system.

The cells of the joint probability tables will show not only symbolic outcomes from Boolean calculations, but will also contain numerical values. These numerical values will help us understand a result obtained from a purely formal manipulation, and *vice versa*. In any case, what we are trying to do here paves the way for the practical applications that must eventually make numerical assignments to abstract probability symbols.

4.2   The Transition Begins

As just alluded to, we are going to start off in a rather abstract way by considering Boolean functions of two variables written as $f(x, y)$. In Chapter One, we learned that Boole's Expansion Theorem allowed us to express Boolean functions as,

$$f(x,y) = f(T,T)(x \circ y) \bullet f(T,F)(x \circ y') \bullet f(F,T)(x' \circ y) \bullet f(F,F)(x' \circ y')$$

This is the way we constructed the full DNF representation for any function $f(x, y)$.

If, as is the case for Classical Logic, we are allowed to assign only T or F to the four coefficients, then we end up with functions like AND when,

$$f(T,T) = T, f(T,F) = F, f(F,T) = F, \text{ and } f(F,F) = F$$

So with this choice, the AND function looks like,

$$
\begin{aligned}
f(x,y) &= f(T,T)(x \circ y) \bullet f(T,F)(x \circ y') \bullet f(F,T)(x' \circ y) \bullet f(F,F)(x' \circ y') \\
&= (T \circ (x \circ y)) \bullet (F \circ (x \circ y')) \bullet (F \circ (x' \circ y)) \bullet (F \circ (x' \circ y')) \\
&= T \circ (x \circ y) \\
&= x \circ y
\end{aligned}
$$

But for a general Boolean function, the carrier set may be defined as, say, $\mathbf{B} = \{a, a', F, T\}$. Then, a function analogous to the classical AND function may be written as $f(x, y) = axy$ where now the coefficient $f(T, T) = a$, and the other three coefficients remain the same as before.

See Figure 4.1 at the top of the next page for the analog to a *joint probability table*. It shows all four terms in the expansion of any Boolean function of two variables. The two columns list x and x', while the two rows list y and y'.

Remember that x stands for $x = T$ and x' stands for $x = F$. The cell at the intersection of the first column and first row (cell 1) is where $x = T, y = T$,

Figure 4.1: *Analog to a joint probability table for the Boolean function $a \circ x \circ y$.*

the cell at the intersection of the second column and first row (cell 2) is where $x = F, y = T$, the cell at the intersection of the first column and second row (cell 3) is where $x = T, y = F$, while the cell at the intersection of the second column and second row (cell 4) is where $x = F, y = F$. Even though the variables x and y *could* assume the values of a or a' in this general function, in both Classical Logic, as well as in probability theory, they are restricted to just T or F.

Thus, this table has laid out the four basis functions appearing in the expansion of any two variable Boolean function. For the beginning example using a function analogous to the classical AND, we know that cells 2, 3, and 4 will contain F, while cell 1 contains,

$$axy = a \circ (T \circ T) = a$$

The *marginal* values of this joint probability table are filled in as well. These marginal values indicate that,

$$(a \bullet F) \bullet (F \bullet F) = T$$

leading to,

$$(a \bullet F) \bullet F = a \bullet F = a = T$$

So, after all of this, we have simplified things down to a T in cell 1 with F in the remaining three cells.

In this case, appealing to a more general Boolean function for the classical AND didn't change anything. We still ended up with T in cell 1 and F in the other three cells. But things get more interesting with analogs to the other classical functions.

Figure 4.1 also illustrates our very first departure from the abstract nature of a Boolean Algebra where no numbers are required. In probability theory, we will use real numbers (as opposed to complex numbers) from 0 to 1 inclusive. And, after we define $P(F) = 0$ and $P(T) = 1$, the joint probability table above can be filled in with cell 1 containing 1 and the remaining three cells containing 0.

Now, I think you can appreciate that this same argument for the binary operator $\wedge$ goes through for the three other binary operators $\diamond$, $\star$, and $\downarrow$ that also assign T to just one term in the function expansion. Their joint probability tables will exhibit the same pattern, with the 1 rotating through cell 1 to cell 2 to cell 3 and finally to cell 4, with 0s in the remaining three cells.

How about any of the six Boolean functions that assign a T to two terms in the expansion? Let's look at one of these Classical Logic functions at the same level of detail as just done above to see if the general argument succeeds. Refer to Figure 4.2 below.

Figure 4.2: *Analog to the joint probability table for the general Boolean function* $(a \circ x \circ y) \bullet (a' \circ x' \circ y')$.

Pick out the binary operator $\leftrightarrow$ called `EQUAL`. Its full DNF expansion is,

$$
\begin{aligned}
f(x, y) &= f(T,T)(x \circ y) \bullet f(T,F)(x \circ y') \bullet f(F,T)(x' \circ y) \bullet f(F,F)(x' \circ y') \\
&= (T \circ (x \circ y)) \bullet (F \circ (x \circ y')) \bullet (F \circ (x' \circ y)) \bullet (T \circ (x' \circ y')) \\
&= (T \circ (x \circ y)) \bullet (T \circ (x' \circ y')) \\
&= (x \circ y) \bullet (x' \circ y')
\end{aligned}
$$

A more general function has a coefficient $f(T,T) = a$ together with a coefficient $f(F,F) = a'$. The other two coefficients remain at F. Thus we now have,

$$
f(x, y) = (a \circ x \circ y) \bullet (a' \circ x' \circ y')
$$

From the expansion, we know that two cells, cell 2 and cell 3, instead of the three cells in the previous example, must contain F. Now, a doesn't reduce to T nor does a' reduce to F. Apparently, we have to keep these abstract elements from the carrier set in cell 1 and cell 4. All the elements in the carrier set are utilized.

A true joint probability table will assign real numbers from 0 to 1 to the cells. Cells 2 and 3 will be set to 0 since they both contain F, while cell 1 with abstract

value a might be set to 0.25, and cell 4 with abstract value a' would then be forced to equal 0.75 as shown in Figure 4.2. Notice that the abstract values must sum to T and the probability assignments must sum to 1.

For the final example, choose one of the four Boolean functions where an F is assigned to only one of the constituent terms in the expansion. Suppose we pick the function IMPLIES, that is, the binary operator $\rightarrow$. Then, following the same line of attack, a new general function analogous to IMPLIES can be written as,

$$f(x, y) = axy + c'x'y + b'x'y'$$

For a change of pace, this function is written like we are used to seeing algebraic equations. The joint probability table is shown in Figure 4.3.

Figure 4.3: *Analog to a joint probability table for the general Boolean function* $(a \circ x \circ y) \bullet (c' \circ x' \circ y) \bullet (b' \circ x' \circ y')$.

The only 0 is placed into cell 3 because this function demands an F when $x = T$ and $y = F$. After all, we wanted to retain the analogy to the Classical Logic meaning of an implication. If the premise is TRUE, but the conclusion is FALSE, logical implication does not hold.

The other three cells have abstract values alleviated by showing some actual legitimate numerical assignments to the cells. However, it is obvious that the carrier set must have been enlarged even more to accommodate all these numerical values. The gory details are left to the exercises.

The important point, though, is that formal manipulations in Boolean Algebra like,

$$b \bullet b' = T$$

and,

$$(a \circ b) \bullet (a \circ b') = a$$

have an analog in probability theory.

4.3   Translating to Logic Functions

Let's do the same thing for the Classical Logic functions as we just did for Boolean functions.[1] Since Classical Logic is just Boolean Algebra with a different notation, no new concepts in formal manipulation are involved. However, the inevitable abstractness inherent in Boolean Algebra is not quite as severe when we make the transition over to logic functions.

For logic functions, as we saw in Chapter Two, the four comparable building blocks for the function expansion of two statements are AB, $\overline{A}B$, $A\overline{B}$, and $\overline{A}\,\overline{B}$. These will correspond to the four cells of the joint probability table.

Let's start by picking a function from the four available where the coefficient function assigns just one T. Since we have already used the example of AND, pick the function NOR this time.

$$f_2(A, B) \equiv A \downarrow B = f(T,T)\,AB \vee f(T,F)\,A\overline{B} \vee f(F,T)\,\overline{A}B \vee f(F,F)\,\overline{A}\,\overline{B}$$

Since $f(F, F) = T$ and the other three coefficient functions assign F, we have,

$$\texttt{Nor[p, q]} \equiv A \downarrow B = \overline{A}\,\overline{B}$$

The joint probability table for this function is shown in Figure 4.4. The 1 has rotated around to cell 4 when compared to the joint probability table for AND.

Figure 4.4: *Joint probability table for the logic function* NOR.

We now consider one of the six logic functions which have two components in their DNF expansion. We looked at EQUAL before, so pick XOR this time.

$$f_9(A, B) \equiv A \oplus B = f(T,T)\,AB \vee f(T,F)\,A\overline{B} \vee f(F,T)\,\overline{A}B \vee f(F,F)\,\overline{A}\,\overline{B}$$

Since $f(T,T)$ and $f(F,F) = F$ while $f(T,F)$ and $f(F,T) = T$, ("exclusive or," one of the two statements is TRUE, but not both of them), we have,

$$\texttt{Xor[p, q]} \equiv A \oplus B = A\overline{B} \vee \overline{A}B$$

---

[1] Douglas Hofstadter builds a fascinating cognitive theory on posing the simple question, What does it really mean to say you are going to do "the same thing?"

A joint probability table for this function is shown in Figure 4.5. The 0s are in cells 1 and 4. A legitimate numerical assignment to cells 2 and 3 is $1/2$. All the entries in the joint probability table must sum to 1, of course, and the marginal probabilities must be correct as well.

	A	$\overline{A}$	
B	0	1/2	1/2
$\overline{B}$	1/2	0	1/2
	1/2	1/2	1

Figure 4.5: *Possible joint probability table for the logic function* XOR.

Finally, we make it easy on ourselves and show the last joint probability table for the function IMPLIES in Figure 4.6. Only the notation for the variables has changed, while the numerical assignments have been kept the same as in Figure 4.3.

	A	$\overline{A}$	
B	0.25	0.10	0.35
$\overline{B}$	0	0.65	0.65
	0.25	0.75	1.00

Figure 4.6: *Possible joint probability table for the logic function* IMPLIES.

The marginal sums for all the joint probability tables work out correctly,

$$P(AB \vee A\overline{B}) = P(A) = 0.25$$

$$P(\overline{A}B \vee \overline{A}\,\overline{B}) = P(\overline{A}) = 0.75$$

$$P(AB \vee \overline{A}B) = P(B) = 0.35$$

$$P(A\overline{B} \vee \overline{A}\,\overline{B}) = P(\overline{B}) = 0.65$$

The numerical total over all four cells is as expected,

$$P(AB \vee \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B}) = P(A \vee \overline{A}) = P(B \vee \overline{B}) = P(\text{TRUE}) = 1$$

Each cell represents a distinct joint statement because conjoining any two cells, for example, $AB \wedge \overline{A}\,\overline{B}$ always results in FALSE. It is important to note that the numerical assignment to each cell of the joint probability tables, that is, to each of these four joint statements,

$$AB, \ \overline{A}B, \ A\overline{B}, \ \text{and} \ \overline{A}\,\overline{B}$$

changed as each new logic function was considered.

4.4   Formal Manipulations for Probabilities

The whole point of demonstrating these formal manipulations in Boolean Algebra and Classical Logic was to bring us to the point where we could list the completely analogous formal manipulations for probabilities. We saw examples in the Boolean Algebra where it was necessary to assign values other than T or F to cells of the joint probability tables for certain functions. A probability wrapped around some statement does the same thing when we want to express a degree of belief that the statement in question is TRUE.

We already have in hand the notation from Classical Logic for joint statements. For example, cell 3 of the joint probability table indexes the joint statement,

$$A\overline{B} \equiv \text{``}A \text{ is TRUE and } B \text{ is FALSE.''}$$

The probability symbol $P(\cdots)$ is wrapped around these joint statements, $P(A\overline{B})$, or their disjunctions, $P(\overline{B}) = P(A\overline{B} \vee \overline{A}\,\overline{B})$ to indicate, by a number from 0 to 1 inclusive, the degree of belief that whatever is inside the parentheses is TRUE.

Repeating the probabilities over the marginal sums,

$$P(AB \vee A\overline{B}) = P(A)$$

$$P(AB \vee \overline{A}B) = P(B)$$

$$P(\overline{A}B \vee \overline{A}\,\overline{B}) = P(\overline{A})$$

$$P(A\overline{B} \vee \overline{A}\,\overline{B}) = P(\overline{B})$$

$$P(AB \vee \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B}) = P(T)$$

we are not then surprised to see these numerical assignments,

$$P(T) = 1$$

$$P(F) = 0$$

$$P(A \vee \overline{A}) = 1$$

$$P(B \vee \overline{B}) = 1$$

$$P(AB \vee \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B}) = 1$$

$$P(\overline{A}) = 1 - P(A)$$

$$P(\overline{B}) = 1 - P(B)$$

If $P(A) = 1$ then $P(\overline{A}) = 0$

The joint statements appearing in the cells of the joint probability table are *mutually exclusive* and *exhaustive*. By mutually exclusive is meant that `AND`ing any two joint statements results in FALSE, or a probability of 0.

Thus, the joint statement in cell 1 "A is TRUE and B is TRUE." may be TRUE or FALSE. A probability between 0 and 1 is then attached to indicate the degree of belief that the joint statement is TRUE, but the joining together of the joint statements in cell 1 and cell 2 "A is TRUE and B is TRUE." `AND` "A is FALSE and B is TRUE." cannot both be TRUE. Therefore, it is FALSE, and we say that these two joint statements are *mutually exclusive*.

By *exhaustive* is meant that all of the joint statements appearing in the joint probability table taken together must be TRUE, or have a probability of 1. The joint statement indexed by cell 1 `OR` the joint statement indexed by cell 2 `OR` the joint statement indexed by cell 3 `OR` the joint statement indexed by cell 4 must be TRUE.

For example, if one of the joint statements is known to be TRUE, then the other three must be FALSE. If two of the joint statements are known to be FALSE, then the other two joint statements must share the total probability and sum to 1. Colloquially, we might summarize all of this by asserting that a degree of belief about any joint statement A and B *must* be placed into *one, and only one*, of the four cells.

4.5   Orthonormal Expansion of Functions

This concept that statements should be *mutually exclusive and exhaustive* is very important in the formal manipulation of probabilities. It arises quite naturally when all the Boolean functions are thought to be ultimately constructed from some set of elementary "building blocks."

Expanding an arbitrary function by relying upon coefficients and some set of fundamental "building blocks" is pervasive throughout all of mathematics. Fourier series and linear algebra are perhaps the most well known areas where there is a heavy conceptual reliance on expanding functions with this in mind.

But we have seen that Boolean Algebra, through Boole's Expansion Theorem, also takes advantage of this notion. And we have exploited this notion in our explanation of the essential meaning of the joint probability table.

Traditionally, function expansion begins with orthonormal functions. Here we have a set of k "building block" functions ϕ_i such that for any two of these functions, $\phi_i \times \phi_j = 0$. This is the "orthogonality" condition between any two functions.

Then this requirement is further augmented by the "normality" condition such that $\sum_{i=1}^{k} \phi_i = 1$. For a Boolean Algebra where notions such as 0, 1, $\times$, and $\sum$ do not exist, the analog orthonormal requirements are that $\phi_i \circ \phi_j = F$ and $\phi_1 \bullet \cdots \bullet \phi_i \cdots \bullet \phi_k = T$.

Here, our set of building block ϕ functions are the $k = 4$ functions AB, $\overline{A}B$, $A\overline{B}$, and $\overline{A}\,\overline{B}$. Any two of these functions are orthogonal since, for example,

$$AB \wedge \overline{A}B = F$$

They satisfy the normality condition because,

$$AB \vee \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B} = T$$

Boole's Expansion Theorem is an example of an orthonormal expansion of a Boolean function. The function $f(A, B)$ is expanded abstractly through,

$$f(A, B) = \sum_{i=1}^{k} c_i(A, B)\, \phi_i(A, B)$$

and, as we have seen, this prescription turns into,

$$f(A, B) = f(T, T)\, AB \vee f(T, F)\, A\overline{B} \vee f(F, T)\, \overline{A}B \vee f(F, F)\, \overline{A}\,\overline{B}$$

The $c_i(A, B)$, the $f(T, T)$ through $f(F, F)$, are called the "coefficients" with respect to the set of basis functions $\phi_i(A, B)$, the AB through $\overline{A}\,\overline{B}$. These building block functions, the basis functions ϕ_i, must stay the same for every different function, while the coefficients will change for every different function. It is just like the situation with different points in a three dimensional space which must have different coordinates, but with the points still defined with respect to the same unchanging set of x, y, and z axes.

4.6   Solved Exercises for Chapter Four

Exercise 4.6.1: Show the operator tables for $\circ$ and $\bullet$ when the carrier set of a Boolean Algebra consists of four elements.

Solution to Exercise 4.6.1

This requires us to go back to Chapter One and collect Tables 1.1 and 1.2, repeated here as Tables 4.1 and 4.2. The introduction to Boolean Algebra in Chapter One, as well as the final CA example in Chapter Three, used four elements in the set **B**. So, even from the beginning, we wanted to emphasize that the functional outcome might be a value other than T or F.

Table 4.1: *The definition of the binary operator " $\circ$ " for a Boolean Algebra with four elements in the carrier set* **B**.

$\circ$	a	a'	F	T
a	a	F	F	a
a'	F	a'	F	a'
F	F	F	F	F
T	a	a'	F	T

Table 4.2: *The definition of the binary operator " $\bullet$ " for a Boolean Algebra with four elements in the carrier set* **B**.

$\bullet$	a	a'	F	T
a	a	T	a	T
a'	T	a'	a'	T
F	a	a'	F	T
T	T	T	T	T

These operator tables were filled in by referring to the axioms that define a Boolean Algebra. Thus, the cells in the table for $x \circ x$, $x \circ x'$, $x \circ F$, and $x \circ T$ were filled in by the **Idempotence, Complementation, and Special Elements axioms**. The **Commutativity axiom** was used to fill in any remaining cells.

Note that in this extension, the `AND` operator in logic is shown as the boxed entries in the four cells of the lower right hand corner of the $\circ$ table. The output is T only if both x and y are T just as in Classical Logic. If either variable, or both, is F, then the operator outputs F just as in Classical Logic.

However, in this extension, there are situations where the operator results in a value other than F or T. For example, $T \circ a$ outputs an a. This may be interpreted as a generalization of the Classical Logic operator AND.

In a similar manner, by looking at Table 4.2 for the $\bullet$ operator, the OR operator in logic is also shown as boxed entries in the lower right hand corner of the table. If either variable, or both, is T, then the $\bullet$ operator outputs a T. Otherwise an F results. There are situations where the operator results in a value other than T or F. For example, $F \bullet a' = a'$ which is the result for the generalization of the logic operator OR.

Let's jump ahead a little bit to discern the motivation for discussing these Boolean generalizations to Classical Logic. With only the two values of T and F available in logic, the AND operator can only return a T or an F given these arguments. With the extension we have just examined, the $\circ$ operator can return a value different than T or F, as in $T \circ a = a$.

This is the same kind of flexibility we want from a probability function where truth values attached to statements are now the arguments. The Classical Logic function XOR can only return a T or an F. If we wrap a probability around two statements A and $\overline{B}$, as in $P(A\,\overline{B})$, then the probability value attached to the joint statement $A\,\overline{B}$ does not have to be exclusively 1 (T) or 0 (F), but can be another value $P(A\,\overline{B}) = a$, where a might be thought of as, say, $1/2$.

Consider another example in our extension where the generic Boolean Algebra result for the $\bullet$ operator is $F \bullet a' = a'$. If we wrap a probability around two statements, as in $P(A \vee B)$, then, again, the probability assignment is not forced to be either 1 (T) or 0 (F), but can be another value $P(A \vee B) = a'$, where a' might be thought of as, say, $1/3$.

Exercise 4.6.2: What might it mean to have the elements in the carrier set "ordered"?

Solution to Exercise 4.6.2

Expand the set **B** so that it now consists of,

$$\mathbf{B} = \{a, a', b, b', F, T\}$$

with four potential values other than T or F. We have the same quintuple as before,

$$(\mathbf{B}, \circ, \bullet, F, T)$$

where the binary operators of $\circ$ and $\bullet$ must continue to obey the axioms of Boolean Algebra. Any variable x can now assume a value of a, a', b, b', as well as the two special elements F and T. The two special elements FALSE and TRUE were the only elements involved in Classical Logic. a' and b' refer, of course, to the elements that are the complements of a and b.

It might be appropriate to introduce the notion of an "ordering relationship" from Boolean Algebra. For example, we could write something like,

$$F \leq a \leq b \leq b' \leq a' \leq T$$

to indicate a relationship among the six elements from the carrier set,

$$\mathbf{B} = \{a, a', b, b', F, T\}$$

The "less than or equal" symbol ($\leq$) is used in an abstract sense to order the elements. Since the elements are abstract entities and not numbers, we can't really use the phrase "less than or equal," but the idea is the same. Eventually, we do want to associate the elements a, a', b, b' with legitimate numerical probability assignments which fall between 0 and 1, as well as including the two anchor points, F and T, corresponding to 0 and 1.

Exercise 4.6.3: Construct an operator table for the binary operator ∘ for the carrier set discussed in the last exercise.

Solution to Exercise 4.6.3

Table 4.3 shows the binary operator table for ∘ given the enlarged set of elements. The operator tables for both ∘ and • will have to grow because of the addition of the new elements in **B**. The ordering relationship imposes new constraints that will help us fill in these tables.

Table 4.3: *Definition of the binary operator " ∘ " for a Boolean Algebra continuing the extension of Classical Logic. The cells not yet defined are marked by a " ⋆ ".*

∘	a	a'	b	b'	F	T
a	a	F	$\star$	$\star$	F	a
a'	F	a'	$\star$	$\star$	F	a'
b	$\star$	$\star$	b	F	F	b
b'	$\star$	$\star$	F	b'	F	b'
F	F	F	F	F	$\boxed{F}$	$\boxed{F}$
T	a	a'	b	b'	$\boxed{F}$	$\boxed{T}$

We have already alluded to the somewhat mechanical procedure for constructing tables such as these. First, the diagonal entries can be filled in immediately because of $x \circ x = x$. The F column can be filled in because of $x \circ F = F$. The T column can be filled in because of $x \circ T = x$. Then the last two rows can be filled in because of $x \circ y = y \circ x$. Any entry for $x \circ x' = F$ can be filled in as well as the commutative entry $x' \circ x = F$.

That leaves the four entries $a \circ b$, $a' \circ b$, $a \circ b'$, and $a' \circ b'$ to be determined with their four corresponding commutative expressions. These are the eight empty cells in Table 4.3 marked with a star ($\star$).

The rules governing the ordering relationship in a Boolean Algebra will fill in the so–far undetermined cells of the $\circ$ operator table. These rules say that,

$$a \circ b \;=\; a$$

$$a \circ b' \;=\; F$$

$$a' \circ b \;=\; b$$

$$a' \circ b' \;=\; b'$$

Thus the fully filled in operator table looks like Table 4.4 below. The one T and three F values make another appearance in the lower right hand corner of the table for the $\circ$ operator. Once again, they are boxed to highlight that this part of the table duplicates the AND operator from Classical Logic.

Table 4.4: *Definition of the binary operator " $\circ$ " for a Boolean Algebra continuing the extension of Classical Logic. Now all the previously empty cells are filled in by referring to the above discussion involving the ordering relationship from Boolean Algebra.*

$\circ$	a	a'	b	b'	F	T
a	a	F	a	F	F	a
a'	F	a'	b	b'	F	a'
b	a	b	b	F	F	b
b'	F	b'	F	b'	F	b'
F	F	F	F	F	$\boxed{F}$	$\boxed{F}$
T	a	a'	b	b'	$\boxed{F}$	$\boxed{T}$

Exercise 4.6.4: Do the same thing for the $\bullet$ operator.

Solution to Exercise 4.6.4

Table 4.5, appearing at the top of the next page, shows the operator table for the binary operator $\bullet$. There are eight missing entries to be filled in for the $\bullet$ operator table, and they are in the same positions as the $\circ$ operator table. We work out the answers after inserting them into their appropriate place in the table.

Table 4.5: *Definition of the binary operator " • " for a Boolean Algebra extending Classical Logic.*

•	a	a'	b	b'	F	T
a	a	T	b	b'	a	T
a'	T	a'	T	a'	a'	T
b	b	T	b	T	b	T
b'	b'	a'	b	b'	b'	T
F	a	a'	b	b'	$\boxed{F}$	$\boxed{T}$
T	T	T	T	T	$\boxed{T}$	$\boxed{T}$

The diagonal entries are filled in by $x \bullet x = x$. The T row can be filled in by $T \bullet x = T$ and the F row with $F \bullet x = x$. Since $x \bullet x' = T$, entries like $a \bullet a' = T$ can be entered. Just as for the $\circ$ table, the four entries $a \bullet b$, $a \bullet b'$, $a' \bullet b$ and $a' \bullet b'$ are filled in from the inclusion rules given above.

$$a \bullet b \;\; = \;\; b$$

$$a \bullet b' \;\; = \;\; b'$$

$$a' \bullet b \;\; = \;\; T$$

$$a' \bullet b' \;\; = \;\; a'$$

Commutativity fills in the other four remaining gaps.

The three T and one F values in the lower right hand corner of the table for the $\bullet$ operator are drawn with a box to highlight that this part of the table duplicates the OR operator from logic. Only in the case that both variables are F will this operation return a F just as in logic. And, if either variable is T, then the operation returns a T just as in Classical Logic.

But in the extension of Classical Logic, there are cases where the operation returns a value that is neither T nor F. These are analogous to the numerical values intermediate between 0 for a FALSE statement and 1 for a TRUE statement. This is exactly the role that probability assumes in the general case.

Exercise 4.6.5: Demonstrate that other axioms are satisfied by these tables.

Solution to Exercise 4.6.5

The **Associativity axiom** says that $x \circ (y \circ z) = (x \circ y) \circ z$. So, for example, should three variables assume the settings of $x = a$, $y = b$, and $z = F$, analyze the left

hand side of the **Associativity axiom** first, followed by the right hand side,

$$x \circ (y \circ z) \;=\; a \circ (b \circ F)$$

$$b \circ F \;=\; F$$

$$a \circ F \;=\; F$$

$$(x \circ y) \circ z \;=\; (a \circ b) \circ F$$

$$a \circ b \;=\; a \qquad\qquad \text{Table 4.4}$$

$$a \circ F \;=\; F$$

$$x \circ (y \circ z) \;=\; (x \circ y) \circ z$$

The **Distributivity axiom** says that,

$$x \bullet (y \circ z) = (x \bullet y) \circ (x \bullet z)$$

Let the variables take on the values of $x = a'$, $y = b'$, and $z = b$.

$$y \circ z \;=\; b' \circ b$$

$$=\; F$$

$$x \bullet (y \circ z) \;=\; a' \bullet F$$

$$=\; a'$$

$$x \bullet y \;=\; a' \bullet b'$$

$$=\; a' \qquad\qquad \text{Table 4.5}$$

$$x \bullet z \;=\; a' \bullet b$$

$$=\; T \qquad\qquad \text{Table 4.5}$$

$$(x \bullet y) \circ (x \bullet z) \;=\; a' \circ T$$

$$=\; a'$$

$$x \bullet (y \circ z) \;=\; (x \bullet y) \circ (x \bullet z)$$

Exercise 4.6.6: Use the formal manipulation rules for Boolean Algebra to illustrate the notion of *mutually exclusive.*

Solution to Exercise 4.6.6

From a probability perspective, we require that the cells in the joint probability table be mutually exclusive. The analogous stipulation within a formal Boolean manipulation context is that the conjunction of any two cells should be F. For example, looking again at the top row of a joint probability table, it is easily worked out that,

$$
\begin{aligned}
(x \circ y) \circ (x' \circ y) &= (x \circ (y \circ x')) \circ y \\
&= (x \circ (x' \circ y)) \circ y \\
&= (x \circ x') \circ (y \circ y) \\
&= F \circ y \\
&= F
\end{aligned}
$$

Thus, from a probability perspective, the joint probability $P(AB \wedge \overline{A}B) = 0$.

Exercise 4.6.7: Illustrate, again using Boolean Algebra, a fact about any marginal sum in a joint probability table.

Solution to Exercise 4.6.7

Another obvious fact from the joint probability table is that $P(AB) \leq P(B)$ and $P(AB) \leq P(A)$. The marginal sum can never be less than one of its components since probabilities are never negative. From the ordering relationship of Boolean Algebra, the analogy would be the Boolean expressions $a \circ b \leq b$ and $a \circ b \leq a$.

Exercise 4.6.8: Show those "gory details" alluded to in this Chapter for the Boolean analog of the `IMPLIES` joint probability table.

Solution to Exercise 4.6.8

Consider once again the 2×2 joint probability tables as shown in Figures 4.3 and 4.6. Let's start off the discussion with a smaller carrier set,

$$\mathbf{B} = \{a, a', b, b', F, T\}$$

to see if we can make things work out. From the joint probability table, we see that the ordering here is,

$$F \leq a \leq b \leq b' \leq a' \leq T$$

The top row of the joint probability table, $P(AB \vee \overline{A}B) = P(B)$, might, through analogy with multiplication of the marginals, be thought of formally in Boolean terms as,

$$(a \circ b) \bullet (a' \circ b) = b \circ (a \bullet a') = b \circ T = b$$

Directly substituting for the two terms from the $\circ$ operator table given in Table 4.4,

$$(a \circ b) \bullet (a' \circ b) = a \bullet b$$

and then from the $\bullet$ operator table given in Table 4.5, confirm this result,

$$a \bullet b = b$$

But we immediately run into a problem. The marginal sum, $P(B) = 0.35$, found by adding 0.25 and 0.10 in the top row of cells, means that b cannot simultaneously mirror the value of both 0.35 and 0.10 from $a' \circ b = b$ from cell 2 of the joint probability table.

We might have foreseen difficulties by looking at the number of possibilities in the joint probability table. There are potentially nine numerical values for the table, consisting of the four individual cell entries together with the four marginal sums and the final total sum of 1. But there are only six elements in $\mathbf{B} = \{a, a', b, b', T, F\}$. We are led to conclude that we must postulate at least one other element and its complement to create an enlarged carrier set.

Let's go ahead and do just that. Add an element c and its complement c' to the carrier set. The ordering of this enlarged set is,

$$F \leq c' \leq a \leq b \leq b' \leq a' \leq c \leq T$$

We will understand how this new ordering comes about in a few moments.

For now, analyze the implications of letting the cells in the joint probability table be filled with the Boolean expressions as shown in Figure 4.7.

Figure 4.7: *Joint probability table with legitimate numerical assignments together with Boolean expressions in the cells and at the margins.*

The first cell, $P(AB)$, has the value a because of $a \circ b = a$. As we learned in the last section, the rules for Boolean ordering dictate such an answer for the $\circ$ operator. Likewise, we achieve our objective for the second cell, $P(\overline{A}B) = c'$, because $a' \circ b = c'$. Here we have taken advantage of the freedom afforded by the enlarged carrier set. The full operator tables for both the $\circ$ and $\bullet$ binary operations will be worked out in later exercises.

Now turn your attention to the various sums across the rows and columns of the joint probability table. These provide further tests and checks on the formal operations. For example, the formal Boolean manipulation rules for the first column yield,

$$(a \circ b) \bullet (a \circ b') \;=\; a \circ (b \bullet b')$$

$$=\; a \circ T$$

$$=\; a$$

which does mirror $P(AB \vee A\overline{B}) = P(A)$. The first transformation above is the **Distributivity axiom** in reverse.

An interesting revelation is obtained by carrying out formal manipulations on the second column of the joint probability table. We could have used the same formal manipulation technique as above to show that,

$$P(\overline{A}B \vee \overline{A}\,\overline{B}) = P(\overline{A})$$

However, if we directly substitute the $\circ$ result for each cell,

$$(a' \circ b) \bullet (a' \circ b') \;=\; c' \bullet b'$$

$$=\; a'$$

we discover a non–obvious relationship that must be satisfied for the $\bullet$ operator table.

This provides an opportunity to mention another one of the standard Boolean axioms, the so–called **De Morgan's axiom**. This axiom will be on display in the next Chapter when we get around to proving the formal aspects of probability symbol manipulation. Here it provides an interesting check on the consistency of the operator tables.

De Morgan's axiom comes in dual forms,

$$(x \bullet y)' \;=\; x' \circ y'$$

$$(x \circ y)' \;=\; x' \bullet y'$$

In the above derivation showing the analogous formal Boolean operations for,

$$P(\overline{A}B \vee \overline{A}\,\overline{B}) = P(\overline{A})$$

everything depended upon $c' \bullet b' = a'$. If this operation is correct, then by **De Morgan's axiom**,

$$c' \bullet b' = (c \circ b)' = a'$$

This, in turn, implies that $b \circ c = a$ within the $\circ$ operator table. These checks on internal consistency are *de rigueur* for any formal system.

One final example from this joint probability table is instructive. From the formal probability manipulation perspective, it is very important to know that $P(A \vee B) = P(A) + P(B) - P(AB)$. This is proven in the next Chapter. For the particular numerical assignments in the current example,

$$P(A \vee B) = 0.25 + 0.35 - 0.25 = 0.35 = P(B)$$

But right now, all we want to do is continue to emphasize that if the formal Boolean operations are to yield the same answer, then the constraints arising from internal consistency must make their appearance in the operator tables. That same answer would correspond to,

$$P(A \vee B) = P(B) = a \bullet b = b$$

If we directly add the marginal probabilities for A and B by forming the Boolean operations that comprise that sum, we find that,

$$
\begin{aligned}
a \bullet b &= (a \circ b) \bullet (a \circ b') \bullet (a \circ b) \bullet (a' \circ b) \\[2mm]
&= (((a \bullet F) \bullet a) \bullet c') \\[2mm]
&= ((a \bullet a) \bullet c') \\[2mm]
&= a \bullet c'
\end{aligned}
$$

Thus, $a \bullet c'$ must equal b to maintain internal consistency. Curiously, we didn't even have to take conscious notice of the fact that the first cell was counted twice. The axioms of Boolean Algebra took care of that for us.

Building on the same technique using **De Morgan's axiom**, we have the further implication for the operator table that,

$$(a \bullet c')' = a' \circ c = b'$$

A later exercise will present a diagram that aids enormously in visualizing these relationships.

Exercise 4.6.9: Prove first that $P(B) = a$ for the joint probability table shown in Figure 4.1, and then that $P(AB) = 1$.

Solution to Exercise 4.6.9

Feel free to intermix the rules for both the Boolean and probability expressions to illustrate any mapping between them. $f(x = T, y = T) \equiv P(AB) = a$. Next,

$f(x = F, y = T) \equiv P(\overline{A}\,B) = F$. This allows us to write,

$$P(B) = P(AB) + P(\overline{A}\,B) = a \bullet F = a$$

But, of course,

$$P(AB) + P(\overline{A}\,B) + P(A\,\overline{B}) + P(\overline{A}\,\overline{B}) = 1$$

or, in the equivalent mapping,

$$a \bullet F \bullet F \bullet F = T$$

So, $P(AB) = a = T = 1$. The Boolean function analog to the AND logic function was $f(x, y) = axy$. That is why from Boole's Expansion Theorem,

$$f(x = T, y = T) \quad = \quad a$$

$$P(AB) \quad = \quad a$$

and,

$$f(x = F, y = T) = F = P(\overline{A}\,B) = 0$$

Exercise 4.6.10: Check that the Associativity axiom is satisfied for the variable settings of $x = b'$, $y = a'$, and $z = T$ by consulting the binary operator table in Table 4.4.

Solution to Exercise 4.6.10

The **Associativity axiom** is,

$$x \circ (y \circ z) = (x \circ y) \circ z$$

The left hand side works out to,

$$x \circ (y \circ z) \quad = \quad b' \circ (a' \circ T)$$

$$= \quad b' \circ a'$$

$$= \quad b'$$

The right hand side works out to,

$$(x \circ y) \circ z \quad = \quad (b' \circ a') \circ T$$

$$= \quad b' \circ T$$

$$= \quad b'$$

Exercise 4.6.11: Is the Distributivity axiom satisfied for the variable settings of $x = b$, $y = F$, and $z = a'$ for the binary operator tables in Tables 4.4 and 4.5?

Solution to Exercise 4.6.11

The **Distributivity axiom** is,

$$x \bullet (y \circ z) = (x \bullet y) \circ (x \bullet z)$$

The left hand side works out to,

$$
\begin{aligned}
x \bullet (y \circ z) &= b \bullet (F \circ a') \\
&= b \bullet F \\
&= b
\end{aligned}
$$

The right hand side works out to,

$$
\begin{aligned}
(x \bullet y) \circ (x \bullet z) &= (b \bullet F) \circ (b \bullet a') \\
&= b \circ T \\
&= b
\end{aligned}
$$

Exercise 4.6.12: Discuss the marginal sum over the second column of Figure 4.2 as both probabilities and as elements from the carrier set.

Solution to Exercise 4.6.12

The marginal sum over the second column of this analog to the joint probability table for the `EQUAL` logic function is $P(\overline{A}) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) = 0 + 0.75 = 0.75$. The Boolean function created to mimic, and, at the same time, generalize this logic function was,

$$f(x, y) = axy \bullet a'x'y'$$

Thus,

$$P(\overline{A}B) = f(x = F, y = T) = F = 0$$

and,

$$P(\overline{A}\,\overline{B}) = f(x = F, y = F) = a'$$

The marginal sum is then,

$$P(\overline{A}) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) = F \bullet a' = a' = 0.75$$

Exercise 4.6.13: Give the full ∘ operator table for understanding the Boolean function that generalizes the IMPLIES joint probability table.

Solution to Exercise 4.6.13

Table 4.6 shows the operator table for the binary operator ∘ with all the entries filled in. The carrier set of this Boolean Algebra has been enlarged to include $\mathbf{B} = \{a, a', b, b', c, c', F, T\}$. The axioms and the ordering relationships of Boolean Algebra will enable one how to figure out the entries in each cell of the operator table. The ordering relationship was given in the text as,

$$F \leq c' \leq a \leq b \leq b' \leq a' \leq c \leq T$$

Table 4.6: *The definition of the binary operator "∘" for a Boolean Algebra that continues the extension of Classical Logic.*

∘	a	a'	b	b'	c	c'	F	T
a	a	F	a	F	a	F	F	a
a'	F	a'	c'	b'	b'	c'	F	a'
b	a	c'	b	F	a	c'	F	b
b'	F	b'	F	b'	b'	F	F	b'
c	a	b'	a	b'	c	F	F	c
c'	F	c'	c'	F	F	c'	F	c'
F	F	F	F	F	F	F	$\boxed{F}$	$\boxed{F}$
T	a	a'	b	b'	c	c'	$\boxed{F}$	$\boxed{T}$

Exercise 4.6.14: Explain how the first row of the ∘ operator table was filled in.

Solution to Exercise 4.6.14

The first two and last two entries are the familiar Boolean operations, for example, $a \circ a = a$ and $a \circ T = a$. $a \circ b = a$ and $a \circ c = a$ come from the ordering relationship where the smaller element is output as a result of the ∘ operation. a is smaller than both b and c. $a \circ b' = F$ and $a \circ c' = F$ again arise from the rules of the Boolean ordering relationship. The ∘ operation with the complement of a larger element results in F. See Exercise 4.6.18 for further intuitive insight into these operations.

Exercise 4.6.15: How might you explain, using the joint probability table and the ordering relationship, that $a' \circ b = c'$ so that the appropriate cell in the second row of the $\circ$ operator table can be filled in.

Solution to Exercise 4.6.15

Looking at the ordering relationship, we see that c is the "largest" value next to T. Now $c \bullet c' = T$ and also decomposes the total probability of 1 for the four cells of the joint probability table. Thus, it's not unreasonable to speculate that c being the largest value might be constituted from, say, three of the four cells in the joint probability table. If these three cells are chosen to be cells 1, 3, and 4,

$$(a \circ b) \bullet (a \circ b') \bullet (a' \circ b') = c$$

Then, we can use standard Boolean operations to reduce this to,

$$
\begin{aligned}
(a \circ b) \bullet (a \circ b') \bullet (a' \circ b') &= (a \circ b) \bullet (b' \circ (a \bullet a')) \\
&= (a \circ b) \bullet (b' \circ T) \\
&= (a \circ b) \bullet b' \\
&= (a \bullet b') \circ (b \bullet b') \\
&= a \bullet b' \\
a \bullet b' &= c
\end{aligned}
$$

Only one cell, cell 2, was left unaccounted for in the above analysis and that was $a' \circ b$ and so it must equal c'. Another way of deriving this is to complement both sides of the result found above and then use **De Morgan's axiom** on the left hand side.

$$(a \bullet b')' = c' \rightarrow a' \circ b = c'$$

Exercise 4.6.16: Give an indirect, but consistent, probabilistic argument that leads to the cell $c \circ b$ in the operator table shown in Table 4.6 being filled in with a.

Solution to Exercise 4.6.16

First of all, by the **Commutativity axiom**, $b \circ c = a$. Next, by **De Morgan's axiom**, $b' \bullet c' = a'$. $P(\overline{A}\,\overline{B}) = a' \circ b' = b'$. $P(\overline{A}B) = a' \circ b = c'$. Finally, we have that $P(\overline{A}\,\overline{B}) + P(\overline{A}B) = b' \bullet c' = a' = P(\overline{A})$.

Exercise 4.6.17: Give the full • operator table for understanding the Boolean function that generalizes the IMPLIES **joint probability table.**

Solution to Exercise 4.6.17

Table 4.7 shows the operator table for • with all the entries filled in.

Table 4.7: *The definition of the binary operator "•" for a Boolean Algebra that continues the extension of Classical Logic.*

•	a	a'	b	b'	c	c'	F	T
a	a	T	b	c	c	b	a	T
a'	T	a'	T	a'	T	a'	a'	T
b	b	T	b	T	T	b	b	T
b'	c	a'	T	b'	c	a'	b'	T
c	c	T	T	c	c	T	c	T
c'	b	a'	b	a'	T	c'	c'	T
F	a	a'	b	b'	c	c'	$\boxed{F}$	$\boxed{T}$
T	T	T	T	T	T	T	$\boxed{T}$	$\boxed{T}$

Exercise 4.6.18: Sketch out a diagram that captures the relationships worked out formally for the last two binary operator tables.

Solution to Exercise 4.6.18

Figure 4.8 at the top of the next page presents a sketch of the ordering relationships that aids in the visualization of the formal properties captured in the ∘ and • operator tables. This kind of visual aid is usually called an "Euler diagram" or a "Venn diagram" and more typically drawn with circles instead of the way I have arranged things.

Exercise 4.6.19: Can the entry of F for $a \circ b'$ be explained with the aid of Figure 4.8?

Solution to Exercise 4.6.19

There is no common overlap in the sections defined by a and b'. Therefore, the entry in the ∘ operator table for $a \circ b'$ is F. $P(A\overline{B})$ is also 0 arising from the $\rightarrow$ logic function as a model assigning numerical values to the joint probability table.
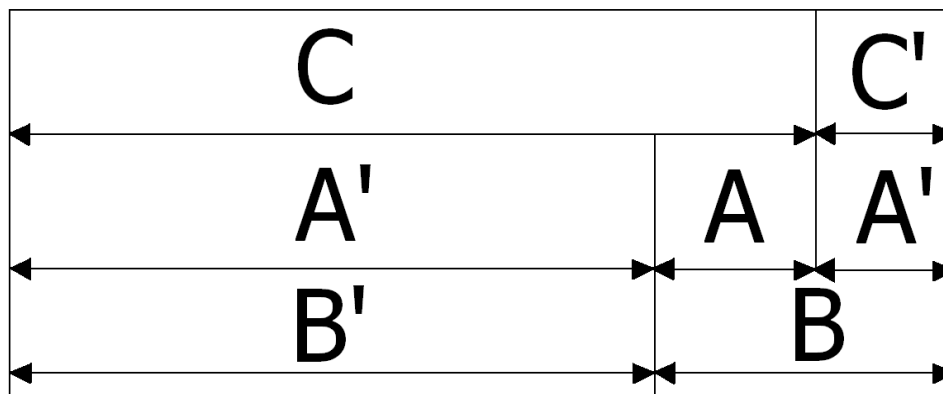
Figure 4.8: *"Euler diagram" that sketches out the ordering relationships for the non–special elements in the carrier set.*

Exercise 4.6.20: Can the entry for $b' \bullet a$ be explained with the aid of Figure 4.8?

Solution to Exercise 4.6.20

Together, the sections defined by b' and a equal the section defined by c. Therefore, the entry in the $\bullet$ operator table for $b' \bullet a$ is c.

Exercise 4.6.21: Use the first version of De Morgan's axiom to double–check the entries in the operator tables for $a \bullet b$ and $a' \circ b'$.

Solution to Exercise 4.6.21

From the first version of **De Morgan's axiom** we know that,

$$(x \bullet y)' = x' \circ y'$$

It was given in the $\bullet$ operation table that $a \bullet b = b$. Thus, $(a \bullet b)' = b'$. But $(a \bullet b)'$ also must equal $a' \circ b'$. Checking the $\circ$ operation table, we find that $a' \circ b'$ does indeed equal b'.

Exercise 4.6.22: Exercise 4.6.20 above used the Euler diagram to try and explain the Boolean operation $b' \bullet a = c$. Use the numerical values in the joint probability table of Figure 4.6 to double–check that $a \bullet b' = c$.

Solution to Exercise 4.6.22

Earlier in Exercise 4.6.15, we showed that,

$$a \bullet b' = c$$

This finding was explained there by showing that $a \bullet b'$ was composed from the Boolean definitions for three cells of the joint probability table.

The numerical values assigned to the first, third, and the last cell of the joint probability table representing $P(AB)$, $P(A\overline{B})$, and $P(\overline{A}\,\overline{B})$ were 0.25, 0, and 0.65. Together they equal 0.90, which, if everything is to be consistent, should be the value for c with then $c' = 0.10$.

Furthermore, $c \bullet c' = T$ and $0.90 + 0.10 = 1.00$. Substitute the answers from the $\circ$ and $\bullet$ operator tables to find that,

$$(a \circ b) \bullet (a \circ b') \bullet (a' \circ b') = (a \bullet F) \bullet b' = a \bullet b' = c$$

Also, check that the Euler diagram shows that together a and b' constitute c.

Chapter 5

Fundamental Rules of Probability

5.1 Introduction

We now arrive at one of the main goals in this introductory Volume. We want to expand our list of important rules for manipulating probabilities. The quintessential nature of these rules is the requirement that they should operate independently of whatever numerical values may have been assigned to the probabilities. We actually got off to a good start in the last Chapter by examining the formal rules of Boolean Algebra in the context of joint probability tables.

A fundamental distinction exists between the formal manipulation of abstract probabilities, and any subsequent assignment of legitimate numerical values to these probabilities. This conceptual distinction cannot be overemphasized. Thus, we are driven to adopt the very rigid viewpoint that an adequate understanding of how an IP makes inferences is enforced through this conceptual division.

Volume I concentrates mainly on the first conceptual notion. Namely, we are concerned with those formal operations that can be carried out without regard to whatever particular numerical values have been attached to the probabilities under some model. That stipulation shouldn't prevent us from illustrating abstract results with numerical examples when such examples are particularly helpful. The equally important second conceptual notion concerning legitimate numerical assignments is treated later in great detail in Volumes II and III.

The previous Chapters have all served as important background material in understanding these formal manipulation rules of probability theory. Therefore, the abstract mathematical structure underlying the manipulation rules for probability will be borrowed from Boolean Algebra.

Despite the inherent abstractness, these rules were easy to understand in the sense that whatever happened was just a repeated mechanical application within a closed finite realm. Proving theorems was essentially a *computational* endeavor, not one, as we tried to emphasize, that demanded any special creative mathematical talent.

It was stressed that in moving towards Information Processing, Classical Logic could be thought of as operating under these same mechanical rules together with an especially simple carrier set consisting of just the two special elements TRUE and FALSE. The generalization to functions of many variables was important, and we showed an interesting application of the formal manipulation rules to cellular automata. An equally important generalization was made to a carrier set with more than two elements.

These generalizations hinted at something that probability theory takes to full fruition. This is the idea that probability theory is a generalization of Classical Logic.

5.2   Notational Refresher

In Chapter Two, a different notation was introduced to distinguish Classical Logic from Boolean Algebra. That notation moves to center stage as we concentrate more and more on probabilities. Simplifications will be invoked as necessary solely in an attempt to keep the syntax from interfering with understanding. Also, we shall refer back to the original notation used in Boolean Algebra whenever that might be helpful or just simpler.

So, once again, let $A, B, C, \cdots$ stand for propositions, that is, statements which are TRUE or FALSE. To indicate an abstract probability, wrap the probability symbol around these statements. Or, more generally, we may wrap a probability around logic functions that have statements as their arguments. For example, we might write expressions like $P(A)$, $P(A \wedge B)$, or later on, $P(B \,|\, A, A \rightarrow B)$.

$P(A)$ represents the degree of belief that statement A is TRUE, $P(A \wedge B)$ represents the degree of belief that the joint statement A and B is TRUE, while the last expression, $P(B \,|\, A, A \rightarrow B)$, represents the degree of belief that B is TRUE *given that* A is TRUE together with the logical implication that A implies B.

In a fully correct notation that takes account of that second fundamental concept mentioned in the **Introduction**, the probability notation should look something like $P(A \,|\, \mathcal{M}_k)$. This is the probability that statement A is TRUE given that some model justifies making a legitimate numerical assignment to the probabilities. This notation stresses that the assignment is conditioned on the *information* represented by some k^{th} model $\mathcal{M}_k$.

Thus, it is a fundamental conceptual notion that there can be NO absolute and unchanging "correct" assignment of a numerical value to a probability.

Depending on what is in $\mathcal{M}_k$, there could be an infinite number of legitimate numerical assignments, with no one of them any more true than another. As a matter of fact, the absolutely essential core feature of a probability is that it MUST change as the information changes.

Nevertheless, for the purpose of deriving formal manipulation rules, we may dispense with explicitly mentioning the conditioning information $\mathcal{M}_k$. It is not strictly necessary when the proviso is attached that we are dealing only with abstract probability symbols.

The rules will be applicable for any and all legitimate assignments made under any conditioning information. Because repetition *ad nauseam* is unfortunately an absolute necessity in trying to drive home these concepts, it must be restated that no numerical probability value can be assigned in a vacuum; there must *always* be some conditioning information, or model, which is the source of that particular numerical assignment.

5.3 Rule for Distributing the Probability Symbol

We have already used the rules, $P(T) = 1$ and $P(F) = 0$. And, in constructing joint probability tables, we observed that a marginal probability was a sum over individual cells, as in, for example, $P(B) = P(AB) + P(\overline{A}B)$.

One of the most important symbolic operations we will use over and over again is to distribute the probability symbol over statements connected by the `OR` operator. For two propositions A and B, the distribution of P over the $\vee$ operator turns out to equal,

$$P(A \vee B) = P(A) + P(B) - P(AB).$$

We can no longer avoid the fact that we have left the comfortable closed finite worlds of Boolean Algebra and Classical Logic. With foreign expressions involving new symbols like $P(\cdots)$, $+$, $-$, 1, and 0, we have entered a new realm. The best we can say is that we have a *mapping* between formal operations in one system and analogous formal operations in a related system.

Consider another typical sketch of a joint probability table for two statements A and B as shown in Figure 5.1. In one formal system, Boolean Algebra, we have,

$$(a \circ b) \bullet (a' \circ b) = b$$

These operations are mapped over to another set of operations in the related system called probability,

$$P(AB) + P(\overline{A}B) = P(B)$$

We find $P(A)$ in the same fashion.

$$(a \circ b) \bullet (a \circ b') \;\; = \;\; a$$

$$P(AB) + P(A\overline{B}) \;\; = \;\; P(A)$$

Figure 5.1: *Joint probability table with symbolic entries representing both Boolean Algebra and probabilities of joint statements in each cell.*

Continue in this same vein by moving back and forth at will between formal operations in the two systems,

$$(a \circ b) \bullet (a' \circ b) \bullet (a \circ b') \bullet (a' \circ b') = T$$

$$P(AB) + P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) = 1$$

$$P(A) = P(AB) + P(A\overline{B})$$

$$P(B) = P(AB) + P(\overline{A}B)$$

$$P(A) + P(B) - P(AB) = P(AB) + P(A\overline{B}) + P(\overline{A}B)$$

$$= (a \circ b) \bullet (a \circ b') \bullet (a' \circ b)$$

$$= a \bullet b$$

$$a \bullet b = P(A \vee B)$$

$$P(A \vee B) = P(A) + P(B) - P(AB)$$

This result can be better appreciated intuitively by referring back to the joint probability table in Figure 5.1. The marginal probability for A, $P(A) = P(AB) + P(A\overline{B})$, (cells 1 and 3), has been added to the marginal probability for B, $P(B) = P(AB) + P(\overline{A}B)$, (cells 1 and 2). But, in this sum, we see that cell 1, $P(AB)$, has been counted twice. Thus, by removing the duplication of cell 1, we have the sums from only cells 1, 2, and 3.

$$P(A \vee B) = P(AB) + P(A\overline{B}) + P(AB) + P(\overline{A}B) - P(AB) = P(AB) + P(\overline{A}B) + P(A\overline{B})$$

5.4   Another Derivation

This derivation of distributing the probability over the $\vee$ operator was an interesting exercise in probability symbol manipulation. Here is another way of justifying the operation of distributing the probability symbol over statements connected by the logical function `OR`.

The purpose of going through another derivation is to provide an opportunity of introducing a very important definition used constantly in the formal manipulation of probabilities,

$$P(A\,|\,B) = \frac{P(AB)}{P(B)}$$

In words, the probability that A is true given that B is true is the probability of the joint statement AB divided by the marginal probability for statement B. Re–expressed as,

$$P(AB) = P(A\,|\,B)\,P(B)$$

it is known as the **Product Rule**.

We'll also need the rule that $P(A) + P(\overline{A}) = 1$ and its obvious consequences in the proof. We call this the **Sum Rule**. We should also highlight the important concept that this rule still applies even when the probability for a statement is conditioned on something else.

Thus, $P(A\,|\,B) + P(\overline{A}\,|\,B)$ still must equal 1, even though the probability for statement A is conditioned upon another statement B. The probability $P(A)$ might be, and in general, will be different than $P(A\,|\,B)$, and $P(\overline{A}\,|\,B)$ might be, and in general, will be different than $P(\overline{A})$. But, in the end, whether statement A is TRUE `OR` statement A is FALSE must be TRUE, no matter how their individual probabilities might change given what they are conditioned on.

This proof shown below closely follows the ones given by Jaynes [12] and Garrett [6]. Unfortunately, it involves a rather longish series of symbol manipulation steps. However, this is by no means atypical when more complicated theorems are derived within some axiom system. Wolfram [20] has some insightful comments on the growth of theorem proofs.

We can begin from where we left off in the derivation at the end of the last section by invoking the **Product Rule** and then the **Sum Rule**.

$$P(A \vee B) \;=\; P(AB) + P(\overline{A}B) + P(A\overline{B})$$

$$P(AB) + P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) \;=\; 1$$

$$P(A \vee B) \;=\; 1 - P(\overline{A}\,\overline{B})$$

$$\;=\; 1 - [P(\overline{A}\,|\,\overline{B})\,P(\overline{B})]$$

$$\;=\; 1 - [\,[\,1 - P(A\,|\,\overline{B})]\,P(\overline{B})\,]$$

$$= \; 1 - [\, P(\overline{B}) - P(A \,|\, \overline{B}) \, P(\overline{B}) \,]$$

$$= \; P(B) + P(A|\overline{B})P(\overline{B})$$

$$= \; P(B) + P(A\overline{B})$$

$$= \; P(B) + P(\overline{B}A)$$

Now continue the derivation by invoking the **Product Rule** once again for $P(\overline{B}A)$ on the right hand side,

$$
\begin{aligned}
P(A \vee B) \; &= \; P(B) + P(\overline{B}A) \\[4pt]
&= \; P(B) + P(\overline{B} \,|\, A) \, P(A) \\[4pt]
&= \; P(B) + [\, [\, 1 - P(B \,|\, A) \,] \, P(A) \,] \\[4pt]
&= \; P(B) + [\, P(A) - P(B \,|\, A) \, P(A) \,] \\[4pt]
&= \; P(B) + P(A) - P(BA) \\[4pt]
&= \; P(A) + P(B) - P(AB)
\end{aligned}
$$

Thus, we have succeeded in proving in another way the important manipulation rule that permits us to distribute the probability operator across the `OR` operator.

5.5   Two Additional Axioms

We didn't present all the axioms for Boolean Algebra in Chapter One. In fact, it is never quite clear in any formal mathematical system what are the minimum number of axioms. And this is true for Boolean Algebra as well.

One could find some minimum pristine set of axioms, and then later derive as theorems what someone else might prefer to directly label as axioms under a more relaxed setting. But the two rules given below are traditionally classified as axioms for Boolean Algebra.

We gained some familiarity with **De Morgan's axiom** in the last Chapter. Both the **Involution axiom** and **De Morgan's axiom** play important roles in derivations of the formal probability rules. One such derivation will follow below after the formal presentation of these two new axioms.

We express these two additional axioms, first in the notation for Boolean Algebra and then in probabilistic notation.

Axiom 7 (Involution)

$$(x')' = x$$

$$P(\overline{\overline{A}}) = P(A)$$

Axiom 8 (De Morgan's Axiom)

$$(x \bullet y)' = x' \circ y'$$

$$(x \circ y)' = x' \bullet y'$$

$$P(\overline{A \vee B}) = P(\overline{A}\,\overline{B})$$

$$P(\overline{AB}) = P(\overline{A} \vee \overline{B})$$

Here is that derivation utilizing these two new axioms to confirm yet another time the distribution of the probability over the $\vee$ operator.

$$P(\overline{A}\,\overline{B}) = P(\overline{A \vee B})$$

$$1 - P(\overline{A \vee B}) = P(\overline{\overline{A \vee B}})$$

$$P(\overline{\overline{A \vee B}}) = P(A \vee B)$$

$$P(A \vee B) = 1 - P(\overline{A}\,\overline{B})$$

$$= P(AB) + P(\overline{A}B) + P(A\overline{B})$$

See Exercise 5.9.20 for another version.

5.6   The Absorption Property

With the very powerful rule for distributing the probability symbol across the OR operator at our disposal, we can continue to prove other elementary properties involved in manipulating probability symbols. Here is a proof of the so–called **Absorption property**.

$$P(A \vee AB) = P(A) + P(AB) - P(AAB)$$

$$P(AAB) = P(AB)$$

$$P(A \vee AB) = P(A) + P(AB) - P(AB)$$

$$= P(A)$$

Here is a proof of another absorption property, dual to the one just given. The theorem just proved is used at the second step.

$$P(A \wedge [\, A \vee B \,]\,) \quad = \quad P(AA \vee AB)$$

$$= \quad P(A \vee AB)$$

$$= \quad P(A)$$

Here is another property that is not immediately self–evident.

$$P(A \vee \overline{A}B) \quad = \quad P(A) + P(\overline{A}B) - P(A\overline{A}B)$$

$$P(A\overline{A}B) \quad = \quad 0$$

$$P(A \vee \overline{A}B) \quad = \quad P(A) + P(\overline{A}B)$$

$$P(\overline{A}B) \quad = \quad P(\overline{A} \,|\, B)\, P(B)$$

$$P(A \vee \overline{A}B) \quad = \quad P(A) + [\, 1 - P(A \,|\, B)\,]\, P(B)$$

$$= \quad P(A) + P(B) - P(A \,|\, B)\, P(B)$$

$$= \quad P(A) + P(B) - P(AB)$$

$$P(A) + P(B) - P(AB) \quad = \quad P(A \vee B)$$

$$P(A \vee \overline{A}B) \quad = \quad P(A \vee B)$$

5.7   The Consensus Property

As a final example for this Chapter of the type of formal manipulations that can be carried out on probabilities, we prove the **Consensus property**.

$$P(AB \vee \overline{A}C \vee BC) = P(AB) + P(\overline{A}C)$$

To accomplish this goal, we once again employ the tactic of using whatever formal Boolean manipulations we desire on the variables, followed by our newly proven rule for distributing the probability symbol across OR.

Let's see what an expansion using the DNF provides. Expand the three terms originally given on the left hand side into six terms,

$$AB \vee \overline{A}C \vee BC = (ABC \vee AB\overline{C}) \vee (\overline{A}BC \vee \overline{A}\,\overline{B}C) \vee (ABC \vee \overline{A}BC)$$

This is the trick of "multiplying" each separate term by an appropriate "1". Thus, for the first term, where $C \vee \overline{C} = T$ plays the role of the 1,

$$(C \vee \overline{C}) \wedge (A \wedge B) = (A \wedge B \wedge C) \vee (A \wedge B \wedge \overline{C})$$

We see that two of these terms, ABC and $\overline{A}BC$, from the six listed in the expansion, are repeated. By repeated invocation of **Commutativity** we can get these two terms together so that,

$$ABC \vee ABC = ABC \text{ as well as } \overline{A}BC \vee \overline{A}BC = \overline{A}BC$$

Dropping these repeated terms we have,

$$AB \vee \overline{A}C \vee BC = (ABC \vee AB\overline{C}) \vee (\overline{A}BC \vee \overline{A}\,\overline{B}C)$$

Factor out AB from the first set of parentheses and $\overline{A}C$ from the second set of parentheses to yield,

$$AB \vee \overline{A}C \vee BC = AB \vee \overline{A}C$$

Since $AB\overline{A}C$ is F, we have no cross term and the probability is simply the sum of the two terms,

$$P(AB \vee \overline{A}C \vee BC) = P(AB) + P(\overline{A}C)$$

If we desire, we could turn this into,

$$P(A|B)\,P(B) + [1 - P(A|C)]\,P(C) = P(C) + P(AB) - P(AC)$$

Dropping the repeated terms was the same as dropping BC,

$$ABC \vee \overline{A}BC = BC(A \vee \overline{A}) = BC$$

now seen as the extraneous term in the original expression.

5.8   Does It Make Sense?

The conclusion to any formal symbolic operation must be subjected to the critique, *"Does it really make any sense?"* For example, here is another probability rule in the same spirit as the ones we have been working on.

$$P(AB) + P(\overline{AB}) = 1$$

One must be careful here about something that has cropped up already in working through the exercises involving the **Involution axiom** and **De Morgan's axiom**. Attention must be paid to the negation notation placed over the statements. The notation $\overline{A} \wedge \overline{B}$ appearing as $\overline{A}\,\overline{B}$ is different than $\overline{A \wedge B}$ appearing as $\overline{AB}$.

Let's see if this result makes sense when we translate $P(AB \vee \overline{AB}) = 1$ into words. "It is certain that one of these two cases must prevail: 1) A and B are both true, or 2) A and B are not both true."

If we decompose the second phrase $\overline{AB}$, alternatively translated as, "It is false that A and B are both true," we see that such a statement further subdivides into three cases: 1) A is false and B is true, or 2) A is true and B is false, or 3) A is false and B is false.

We have harped on the fact that the four cells of the 2×2 joint probability table are mutually exclusive and exhaustive, which is to say,

$$P(AB \vee \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B}) = P(AB) + P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) = 1$$

Using the familiar Boolean operations on the variables A and B, we can integrate both the verbal expressions and the joint probability table into a meaningful whole. The exercise also provides an opportunity to carefully distinguish the placement of the negation signs over the variables. We first expand $\overline{A}$ and $\overline{B}$ to,

$$\overline{A} \;=\; \overline{A}B \vee \overline{A}\,\overline{B}$$

$$\overline{B} \;=\; A\overline{B} \vee \overline{A}\,\overline{B}$$

so that we can now write,

$$\overline{A} \vee \overline{B} \;=\; \overline{A}B \vee \overline{A}\,\overline{B} \vee A\overline{B} \vee \overline{A}\,\overline{B}$$

$$\;=\; \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B}$$

By **De Morgan's axiom**,
$$\overline{A \wedge B} = \overline{A} \vee \overline{B}$$

We now have what we want because,

$$\overline{AB} = \overline{A}B \vee A\overline{B} \vee \overline{A}\,\overline{B}$$

The three terms on the right hand side, when translated into words, are exactly the three cases listed above for, "It is false that A and B are both true." These three terms are also cells 2, 3, and 4 of the joint probability table, thus permitting us to write,

$$P(AB) + P(\overline{AB}) = 1$$

We might write the following equation,

$$P(\overline{A}\,\overline{B}) + P(\overline{\overline{A}\,\overline{B}}) = 1$$

with a quick appeal to the **Sum Rule**. We can appreciate this equation in a new light through the arguments we have just presented. The easiest would be to refer back to the joint probability table, and claim that the first term, $P(\overline{A}\,\overline{B})$, references cell 4, while the overbar in the second term references everything else, that is, cells 1, 2, and 3. And together, cells 1, 2, 3, and 4 in the joint probability table certainly do add up to 1.

5.9   Solved Exercises for Chapter Five

Exercise 5.9.1: List some manipulation rules which arise from the axioms, lemmas, and theorems for Boolean Algebra as proved in Chapter One.

Solution to Exercise 5.9.1

$$
\begin{aligned}
P(AA) &= P(A) & &\textbf{Idempotence} \\
P(A \vee A) &= P(A) & &\textbf{Idempotence} \\
P(A \vee \overline{A}) &= 1 & &\textbf{Complementation} \\
P(A\overline{A}) &= 0 & &\textbf{Complementation} \\
P(AAB) &= P(AB) & &\textbf{Theorem 1} \\
P(CCCBBA) &= P(ABC) & &\textbf{Theorem 2} \\
P(ABC\overline{A}) &= 0 & &\textbf{Theorem 3}
\end{aligned}
$$

Exercise 5.9.2: How we can begin to think about probability theory generalizing Classical Logic from these rules?

Solution to Exercise 5.9.2

Many texts reasonably assert that logic originated in what was found acceptable as an argument in debate or rhetoric. Gradually, this attitude transitioned into what was considered a correct use of language. Thus, we can imagine that the Greeks fully accepted an argument in this language, "Either Socrates is a man or Socrates is not a man." Probability generalizes this with the rule, $P(A \vee \overline{A}) = 1$, which, in a rather stilted manner, says that "the degree of belief concerning A or not A being true must be a certainty."

Similarly, one can imagine that a rhetorical argument in this language was the hallmark of proper reasoning, "It cannot be true that Socrates is both a man and not a man at the same time." Once again, probability theory reproduces the same argument with $P(A \wedge \overline{A}) = 0$, with the interpretation, "The degree of belief that A and not A cannot both be simultaneously true is a certainty."

Exercise 5.9.3: Using the notation for variables and operations from Boolean Algebra, what is $P(BAB)$?

Solution to Exercise 5.9.3

Repeating Lemma 4,

$$
\begin{aligned}
(y \circ x) \circ y &= (x \circ y) \circ y \\
&= x \circ (y \circ y) \\
&= x \circ y
\end{aligned}
$$

Therefore, $P(BAB) = P(AB)$.

Exercise 5.9.4: Using a theorem from Chapter One, what is $P(BBAABBB)$?

Solution to Exercise 5.9.4

Theorem 1 showed, just as in the previous exercise, that the pattern of Bs and As could be reduced to simply to AB. Thus, $P(BBAABBB) = P(AB)$.

Exercise 5.9.5: Using the notation for variables and operations from Boolean Algebra, what is $P(A\,\overline{B}\,\overline{A}\,B)$?

Solution to Exercise 5.9.5

This is an example of why the cells in a joint probability table are mutually exclusive.

$$
\begin{aligned}
(x \circ y') \circ (x' \circ y) &= x \circ (y' \circ x') \circ y \\
&= x \circ (x' \circ y') \circ y \\
&= (x \circ x') \circ (y' \circ y) \\
&= F \circ F \\
&= F
\end{aligned}
$$

Therefore, $P(A\overline{B}\,\overline{A}B) = P(F) = 0$.

Exercise 5.9.6: What is the probability in cell 2 and cell 3 of the generic 2 × 2 joint probability table?

Solution to Exercise 5.9.6

Cell 2 indexes the joint statement $\overline{A}B$ and cell 3 indexes the joint statement $A\overline{B}$. Thus,

$$P(\overline{A}B \vee A\overline{B}) = P(\overline{A}B) + P(A\overline{B}) - P(\overline{A}BA\overline{B})$$

The previous exercise showed that $P(\overline{A}BA\overline{B})$ was 0. Therefore, these two cells are mutually exclusive, and,

$$P(\overline{A}B \vee A\overline{B}) = P(A\overline{B}) + P(\overline{A}B)$$

Exercise 5.9.7: Revisit De Morgan's axiom with expressions in Boolean Algebra notation together with the probability formulas.

Solution to Exercise 5.9.7

$$P(\overline{A \vee B}) = (x \bullet y)' = x' \circ y' = P(\overline{A} \wedge \overline{B})$$

$$P(\overline{A \wedge B}) = (x \circ y)' = x' \bullet y' = P(\overline{A} \vee \overline{B})$$

Exercise 5.9.8: Using the assistance of a 2 × 2 joint probability table for two variables A and B, discuss the Sum Rule as used in section 5.8.

Solution to Exercise 5.9.8

The **Sum Rule** says that,

$$P(AB) + P(\overline{AB}) = 1$$

The marginal sum for $P(\overline{A})$ is $P(\overline{A}B) + P(\overline{A}\,\overline{B})$. Likewise, the marginal sum for $P(\overline{B})$ is $P(A\overline{B}) + P(\overline{A}\,\overline{B})$. Adding these two marginal sums, we see that $P(\overline{A}\,\overline{B})$ has been counted twice.

$$P(\overline{A}) + P(\overline{B}) = P(\overline{A}\,B) + P(\overline{A}\,\overline{B}) + P(A\,\overline{B}) + P(\overline{A}\,\overline{B})$$

Therefore, if we subtract $P(\overline{A}\,\overline{B})$ from the sum we have the result,

$$P(\overline{A}) + P(\overline{B}) - P(\overline{A}\,\overline{B}) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(A\overline{B})$$

But the left hand side of the above equation is the result of applying the rule for distributing the probability symbol across the $\vee$ operator,

$$P(\overline{A} \vee \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A}\,\overline{B})$$

Axiom 8, **De Morgan's axiom**, lets us assert that,

$$P(\overline{AB}) = P(\overline{A} \vee \overline{B})$$

resulting in,

$$P(\overline{AB}) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(A\overline{B})$$

From the requirement that all four cells of the joint probability table must be exhaustive, we know that,

$$P(AB) + P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) = 1$$

confirming that,

$$1 - P(AB) = P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) = P(\overline{AB})$$

An easy way to double–check all of this is to revisit the original expression,

$$P(\overline{A}) + P(\overline{B}) - P(\overline{A}\,\overline{B})$$

The joint probability table tells us that, in terms of cell numbers, this is,

$$2 + 4 + 3 + 4 - 4 = 2 + 3 + 4$$

The probabilities in cells 2, 3, and 4 are, in fact,

$$1 - P(AB) = P(\overline{AB})$$

Exercise 5.9.9: Generalize these results found in the last exercise to three statements.

Solution to Exercise 5.9.9

The comparable result for variables A, B, and C would be,

$$P(ABC) = 1 - P(\overline{ABC})$$

Imagine a $2 \times 2 \times 2$ joint probability table with eight cells. $P(\overline{ABC})$ consists of the seven joint probabilities like $P(A\overline{B}C)$, $P(A\overline{B}\,\overline{C})$, $\cdots$, $P(\overline{A}\,\overline{B}\,\overline{C})$ situated in the last seven cells of this joint probability table with $P(ABC)$ in the first cell. In words, $\overline{ABC}$ is "It is false that A, B, and C are all simultaneously true." $P(\overline{ABC})$ is the *state of knowledge* represented by the real number between 0 and 1 that this joint statement, "It is false that A, B, and C are all simultaneously true.", is true.

Exercise 5.9.10: Using the logic functions as described in Chapter Two, prove De Morgan's axiom.

Solution to Exercise 5.9.10

Consider first this version of **De Morgan's axiom** in logic notation,

$$\overline{A \vee B} = \overline{A} \wedge \overline{B}.$$

The OR operator is $f_{15}(A, B)$ which has the functional assignments of $TTTF$ for all four variable settings. Then apply the NOT A operator on $A \vee B$ as the first argument as in, $f_6 [f_{15}(A, B), B]$. The output for all four possible variable settings is $FFFT$. This result is the same as,

$$\overline{A} \wedge \overline{B} \equiv f_5 [f_6(A, B), f_7(A, B)] = FFFT$$

Therefore, the two expressions are logically equivalent, $(\overline{A \vee B}) \leftrightarrow (\overline{A} \wedge B)$ is T for all variable settings.

Now consider the second version of **De Morgan's axiom**,

$$\overline{A \wedge B} = \overline{A} \vee \overline{B}$$

The AND operator is $f_5(A, B)$ which has the functional assignments of $TFFF$ for all four variable settings. The NOT A operator on this result, $f_6 [f_5(A, B), B]$, is $FTTT$. The right hand side is $f_{15} [f_6(A, B), f_7(A, B)]$ which also works out to $FTTT$ for all four variable settings. The left and right hand sides are equal for all variable settings, therefore **De Morgan's axiom** does represent a tautology or a logical equivalence.

Exercise 5.9.11: Use the Sum and Product Rule to expand $P(\overline{A}\,\overline{B})$ and then show that the result is easily explained by referring back to the 2×2 joint probability table.

Solution to Exercise 5.9.11

The **Product Rule** is applied first, followed by the **Sum Rule**. The **Product Rule** is applied again in reverse fashion at the last step.

$$
\begin{aligned}
P(\overline{A}\,\overline{B}) &= P(\overline{A} \,|\, \overline{B}) \, P(\overline{B}) \\[2mm]
&= [1 - P(A \,|\, \overline{B})] \, P(\overline{B}) \\[2mm]
&= P(\overline{B}) - [P(A \,|\, \overline{B}) \, P(\overline{B})] \\[2mm]
&= P(\overline{B}) - P(A\overline{B})
\end{aligned}
$$

$P(\overline{B})$ is the marginal sum over cells 3 and 4. Subtracting $P(A\overline{B})$, Cell 3, leaves just Cell 4, which is $P(\overline{A}\,\overline{B})$.

Exercise 5.9.12: Generalize the distribution of P across the OR function of three variables. Use the $2 \times 2 \times 2$ joint probability table for a heuristic justification.

Solution to Exercise 5.9.12

We would like to find the generalization for $P(A \vee B \vee C)$. **De Morgan's axiom** generalizes to more than two variables. Thus, $\overline{A \vee B \vee C} = \overline{A} \wedge \overline{B} \wedge \overline{C}$.

Invoking the **Involution axiom** just as we did for the two variable case,

$$A \vee B \vee C = \overline{\overline{A} \wedge \overline{B} \wedge \overline{C}}$$

We now have,

$$P(A \vee B \vee C) = 1 - P(\overline{A} \wedge \overline{B} \wedge \overline{C})$$

Imagine that the $2 \times 2 \times 2$ joint probability table has been constructed as depicted below in Figure 5.2. Of course, 1 is the sum of the probabilities in all eight cells.

Figure 5.2: *Joint probability table for three variables A, B, and C.*

$P(\overline{A}\,\overline{B}\,\overline{C})$ is in cell 8, so $P(A \vee B \vee C)$ must be the sum over the first seven cells of the $2 \times 2 \times 2$ joint probability table. The usual way this result is presented is,

$$P(A \vee B \vee C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

This is the same as above if we substitute in the appropriate cells for the indicated marginal probabilities. Thus,

$$
\begin{aligned}
P(A) &= \text{Cells } 1+2+3+4 \\[6pt]
P(B) &= \text{Cells } 1+3+5+7 \\[6pt]
P(C) &= \text{Cells } 1+2+5+6 \\[6pt]
P(AB) &= \text{Cells } 1+3 \\[6pt]
P(AC) &= \text{Cells } 1+2 \\[6pt]
P(BC) &= \text{Cells } 1+5 \\[6pt]
P(ABC) &= \text{Cell } 1
\end{aligned}
$$

The three subtraction terms take away one of the double occurrences for cells 2, 3, and 5. These terms also take away any presence of cell 1, so the addition of $P(ABC)$ as the final term throws cell 1 back to the mix. Totaling up all seven terms, we have the addition of cells 1 through 7, in other words, just the probability that is required.

Exercise 5.9.13: Use the Product and Sum Rules on De Morgan's axiom expansion of $P(A \vee B \vee C)$ for another derivation on how to distribute the P symbol.

Solution to Exercise 5.9.13

This is a completely circular argument, but at least it is practice is applying the **Sum** and **Product** rules,

$$
\begin{aligned}
P(A \vee B \vee C) &= P(\overline{\overline{A} \wedge \overline{B} \wedge \overline{C}}) \\[2mm]
&= 1 - P(\overline{A} \wedge \overline{B} \wedge \overline{C}) \\[2mm]
&= 1 - [P(\overline{A} \,|\, \overline{B}\,\overline{C})\, P(\overline{B}\,|\,\overline{C})\, P(\overline{C})] \\[2mm]
&= 1 - [1 - P(A\,|\,\overline{B}\,\overline{C})]\, P(\overline{B}\,|\,\overline{C})\, P(\overline{C})]] \\[2mm]
&= 1 - P(\overline{B}\,|\,\overline{C})\, P(\overline{C}) + P(A\,|\,\overline{B}\,\overline{C})\, P(\overline{B}\,|\,\overline{C})\, P(\overline{C}) \\[2mm]
&= 1 - P(\overline{B}\,\overline{C}) + P(A\overline{B}\,\overline{C}) \\[2mm]
&= 1 - [\, P(A\overline{B}\,\overline{C}) + P(\overline{A}\,\overline{B}\,\overline{C})\,] + P(A\overline{B}\,\overline{C}) \\[2mm]
&= 1 + P(A\overline{B}\,\overline{C}) - P(A\overline{B}\,\overline{C}) - P(\overline{A}\,\overline{B}\,\overline{C}) \\[2mm]
&= 1 - P(\overline{A}\,\overline{B}\,\overline{C})
\end{aligned}
$$

This confirms our findings for $P(A \vee B \vee C)$ from the previous exercise. Or even more succinctly,

$$[\,(x \bullet y) \bullet z\,]' = (x' \circ y') \circ z' \rightarrow 1 - P(A \vee B \vee C) = P(\overline{A}\,\overline{B}\,\overline{C}) \rightarrow P(A \vee B \vee C) = 1 - P(\overline{A}\,\overline{B}\,\overline{C})$$

Exercise 5.9.14: Verify the Consensus property.

Solution to Exercise 5.9.14

Near the end of section 5.7, we purported to show through formal manipulations that,

$$P(C) + P(AB) - P(AC) = P(AB) + P(\overline{A}C)$$

The relevant marginal probabilities from the joint probability table in Figure 5.2 are,

$$P(C) \quad = \quad \text{Cells } 1+2+5+6$$

$$P(AB) \quad = \quad \text{Cells } 1+3$$

$$P(AC) \quad = \quad \text{Cells } 1+2$$

$$P(\overline{A}C) \quad = \quad \text{Cells } 5+6$$

The left hand side consists of cells $1+2+5+6$ plus cells $1+3$ minus cells $1+2$, thus yielding cells $1+3+5+6$. The right hand side is cells $1+3$ plus cells $5+6$, thus yielding cells $1+3+5+6$. The identity has been confirmed.

Exercise 5.9.15: Using the binary operations on logic functions from Chapter Two, verify the logical equivalence between the left and right sides of the Consensus property for at least one possible setting of the variables.

Solution to Exercise 5.9.15

Logical equivalence is the same as showing that the `EQUAL` operator returns the value T for its left and right hand side arguments. In other words, the logic expression

$$(AB \vee \overline{A}C \vee BC) \leftrightarrow (AB \vee \overline{A}C)$$

should return T for all eight possible variable assignments.

There are three variables A, B, and C involved in this logical equivalency, and therefore eight possible values that these three variables could assume. Arbitrarily, assign $A = T$, $B = F$, and $C = T$ as one of these eight possibilities, and solve for the logic function as given above.

Start first with the logic expression on the left hand side, $AB \vee \overline{A}C \vee BC$, and substitute in the particular choice for the variables as selected above.

$$AB \vee \overline{A}C \vee BC \quad = \quad (T \wedge F) \vee (F \wedge T) \vee (F \wedge T)$$

$$= \quad F \vee F \vee F$$

$$= \quad F$$

Now the right hand side,

$$(AB \vee \overline{A}C) \quad = \quad (T \wedge F) \vee (F \wedge T)$$

$$= \quad F \vee F$$

$$= \quad F$$

The left hand expression evaluates to F and the right hand expression also evaluates to F. Therefore, $F \leftrightarrow F$, that is, $f_8(F, F)$, returns T.

Appendix D presents a *Mathematica* program that does all the hard work of checking all eight combinations. This program will therefore demonstrate that a logical equivalency does exist for the **Consensus property**. That is, the logic function returns T for all eight possible settings as we just proved it returned T for one of those possibilities.

Exercise 5.9.16: Show how De Morgan's axiom explains the "duality" relating $a \circ b'$ and $a' \bullet b$ in Tables 4.4 and 4.5.

Solution to Exercise 5.9.16

When the ordering relationship $a \leq b$ is imposed, the binary operation tables show that,

$$a \circ b' \;=\; F$$

$$a' \bullet b \;=\; T$$

Using **De Morgan's axiom**,

$$(a \circ b')' = F' \overset{\text{Duality}}{\Longrightarrow} a' \bullet b = T$$

Thus, executing a "duality" operation means that each variable is complemented, the binary operators are exchanged, and the special elements T and F are also reversed.

Exercise 5.9.17: How might you interpret this last result in terms of probabilities? Use the 2×2 joint probability table.

Solution to Exercise 5.9.17

Let $a \circ b'$ be the analog to $P(A\overline{B})$. Cell 3 in the 2×2 joint probability table contains $P(A\overline{B})$ which is equal to 0 since $a \circ b' = F$. The analog to $(a \circ b')' = T$ is $1 - P(A\overline{B}) = 1$.

Then, since all four cells of the joint probability table must sum to 1, and $P(A\overline{B}) = 0$, we have,

$$P(AB) + P(\overline{A}B) + P(\overline{A}\,\overline{B}) = 1$$

But the analog to $a' \bullet b$ is,

$$P(\overline{A} \vee B) = P(\overline{A}) + P(B) - P(\overline{A}B)$$

from the theorem on distributing P over the $\vee$ operator. Forming the marginal sums,

$$P(\overline{A}) = P(\overline{A}B) + P(\overline{A}\,\overline{B})$$

and

$$P(B) = P(AB) + P(\overline{A}B).$$

Subtracting $P(\overline{A}B)$ once, we have that $a' \bullet b = T$ is the same as,

$$P(\overline{A} \vee B) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(AB) = 1$$

and the analogy with the exercise above is complete.

Exercise 5.9.18: Give a probability slant to the XOR logic function $A \oplus B$.

Solution to Exercise 5.9.18

From the definition of the XOR function, F is assigned if A and B are both T or both F. T is assigned in the other two cases when A is T and B is F or A is F and B is T. This is the origin of our verbal understanding of "exclusive or," that is, one of the two variables is true, but not both.

From this definition, we also derived the DNF expansion,

$$A \oplus B = \overline{A}B \vee A\overline{B}$$

Thus,

$$P(\overline{A}B \vee A\overline{B}) = P(\overline{A}B) + P(A\overline{B})$$

Cells 1 and 4 must have a 0 entered because these are the conditions where the functional assignment is F. Cells 2 and 3 could contain any legitimate numerical value such that,

$$P(\overline{A}B) + P(A\overline{B}) = 1$$

In upcoming Chapters, we will examine situations expressed similar to,

$$P(B \,|\, A, \mathcal{M}_k \equiv \text{XOR})$$

What is an IP's degree of belief about the statement B if XOR is given as the model? We are also given that A is true as well. B could not be true along with the supposition that A is true and XOR is true. Therefore, $\overline{B}$ must be true, and consequently,

$$P(\overline{B} \,|\, A, \mathcal{M}_k \equiv \text{XOR}) = 1 \text{ and } P(B \,|\, A, \mathcal{M}_k \equiv \text{XOR}) = 0$$

As the formal rules for probability manipulation continue to develop, they will confirm this result.

Exercise 5.9.19: Using the formal manipulation rules of Boolean Algebra, derive the Absorption property.

Solution to Exercise 5.9.19

Arguing using the newly found formal rules for probability manipulation, we found that,

$$P(A \vee \overline{A}B) = P(A \vee B)$$

But the Absorption property is perhaps more easily seen working on the analogous Boolean expression,

$$
\begin{aligned}
x \bullet (x' \circ y) &= (x \bullet x') \circ (x \bullet y) \\[2mm]
&= T \circ (x \bullet y) \\[2mm]
&= (T \circ x) \bullet (T \circ y) \\[2mm]
&= x \bullet y
\end{aligned}
$$

Exercise 5.9.20: Using the formal manipulation rules of Boolean Algebra, derive $P(A \vee B)$.

Solution to Exercise 5.9.20

One would agree that what follows is a rather curious derivation,

$$x \bullet y \equiv P(A \vee B)$$

$$(x \bullet y) \circ T \equiv P(A \vee B) \times 1$$

$$(y \bullet x) \circ (y \bullet y') \equiv P[\,(B \vee A) \wedge (B \vee \overline{B})\,]$$

$$y \bullet (x \circ y') \equiv P(B) + P(A\overline{B})$$

$$(y \circ T) \bullet (x \circ y') \equiv [\,P(B) \times 1\,] + P(A\overline{B})$$

$$(y \circ (x \bullet x')) \bullet (x \circ y') \equiv P[\,(B \wedge (A \vee \overline{A})\,] + P(A\overline{B})$$

$$(x \circ y) \bullet (x' \circ y) \bullet (x \circ y') \equiv P(AB) + P(\overline{A}B) + P(A\overline{B})$$

Chapter 6

Bayes's Theorem

6.1 Introduction

We devote a separate Chapter to introduce that most famous of all probability manipulation rules. Bayes's Theorem is universally invoked when making inferences. It is actually quite easy to derive in comparison with some of the other symbolic exercises we have already experienced.

Perhaps the most important thing to keep in mind about Bayes's Theorem is exactly what its name implies. It is a *mathematical theorem*. Therefore, it should be believed as much, and under the same circumstances, as we believe in any mathematical theorem.

When thought about in a formal abstract sense, Bayes's Theorem should never have acquired the sobriquet of being *controversial*. It is simply on a par with all of the other theorems involving probability manipulations that we have been proving along the way.

It is a rule that is applicable to probabilities in general, irrespective of what particular numerical assignment might have been made to the probabilities. The numerical assignments belong to their own conceptual realm, separate from the formal manipulation rules.

The manipulation rule that is called Bayes's Theorem comes into its own when we seek to generalize Classical Logic within probability theory. In the next Chapter, we will begin the all–important process of showing how probability can reproduce any result from Classical Logic. With this realization, we then proceed beyond Classical Logic as *deduction* to the techniques for making *inferences*.

We turn now to how different versions of Bayes's Theorem may be written down. Deriving these alternative expressions relies upon the formal manipulation rules for probabilities as built up over the past few Chapters.

6.2   Different Versions

In section 5.4, the **Product Rule** allowed us to write the probability expression,

$$P(A \mid B) = \frac{P(AB)}{P(B)} \tag{6.1}$$

This is the simplest version of Bayes's Theorem.

Since $P(B)$ is a marginal probability, it may be expanded into the sum of its constituent joint probabilities,

$$P(B) = P(AB) + P(\overline{A}B)$$

The second version of Bayes's Theorem looks like this,

$$P(A \mid B) = \frac{P(AB)}{P(AB) + P(\overline{A}B)} \tag{6.2}$$

It is not the relative ease with which these expressions are derived, but rather their consequences that we wish to emphasize. For example, $P(AB)$ must be less than, or possibly equal to, $P(B)$. Since both the numerator and denominator in the right hand side of Bayes's Theorem are real numbers between 0 and 1, $P(A \mid B)$ on the left hand side must remain a real number between 0 and 1.

In the case that $P(AB)$ is equal to $P(B)$, then $P(A \mid B)$ is equal to 1. In the case that $P(AB)$ is equal to 0, then $P(A \mid B)$ is equal to 0. In all other cases, $P(AB)$ is less than $P(B)$, so $P(A \mid B)$ remains between 1 and 0.

Derive now a third version of Bayes's Theorem by invoking the **Commutativity axiom** and the **Product Rule** on Equation (6.2),

$$P(AB) \;=\; P(BA)$$

$$P(BA) \;=\; P(B \mid A)\, P(A)$$

$$P(\overline{A}B) \;=\; P(B\overline{A})$$

$$P(B\overline{A}) \;=\; P(B \mid \overline{A})\, P(\overline{A})$$

leading to,

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B \mid A)\, P(A) + P(B \mid \overline{A})\, P(\overline{A})} \tag{6.3}$$

In any version of Bayes's Theorem, we have left the comfortable closed world of a Boolean Algebra. Most noticeably, we have a division of two real numbers. Neither numbers nor division are part of a Boolean Algebra. Division by zero is forbidden in arithmetic, and so we have the consequence that $P(B)$ cannot equal 0.

Recall any of the four logic functions, `FALSE`, `NOR`, `DIFFERENCE`, or `NOT B`. These were functions f_1, f_2, f_4, and f_7 with the notation $\perp$, $\downarrow$, $\diamond$, and $\dashv$ if the syntax emphasized their role as binary operators. All four of these functions had a DNF expansion where two coefficients were $f(T, T) = F$ and $f(F, T) = F$. In other words, for the two possible cases where $B = T$, these functions output an F. The probability that statement B was TRUE was always $P(B) = 0$.

Therefore, it is impossible to condition on a statement asserting that B is TRUE when the model asserts it never will be TRUE. Conditioning on a contradiction is never allowed in Classical Logic. Appropriately, it should never be allowed in probability theory either.

As an example, computation of the expression $P(A = T \mid B = T, \text{NOT B})$ is not permitted under Bayes's Theorem because under the model of the `NOT B` logic function the numerical assignment to $P(B)$ is zero.

On the other hand, placing statement B to the right of the "|" symbol in $P(A \mid B)$ does NOT mean that $P(B) = 1$, or that B MUST HAPPEN. If that were the case, then $P(A \mid B)$ would always equal $P(AB)$.

By placing the statement B in this position, we indicate that it is not always FALSE under the model, and an assignment of positive probability MAY HAPPEN under some model. If B does happen, then revise the probability for A by Bayes's Theorem. If B CANNOT HAPPEN under any circumstances whatsoever, then we are back to the logical contradictory position of asserting that something might be true when it never can be true.

This notion is less jarring when discussing the probability for some particular set of data within the context of Bayes's Theorem. When some statement A is conditioned on statement B, here the occurrence of the actual data, we don't think that these data had to happen, only that they could have happened. Other outcomes for the data are accepted as possibilities, and, if at some other time, these other outcomes take place, then they are processed by Bayes's Theorem in the same manner as the first set of data.

6.3   Generalizing to Additional Variables

It is a pleasant feature that the introductory version of Bayes's Theorem shown as Equation (6.1) can be generalized in the expected manner. Suppose we want the manipulation rule that tells us how to update the uncertainty surrounding statement A when conditioned on the truth of statements B and C. Bayes's Theorem is then written as,

$$P(A \mid BC) = \frac{P(ABC)}{P(BC)} \tag{6.4}$$

If we want to expand the denominator we have,

$$P(A \mid BC) = \frac{P(ABC)}{P(ABC) + P(\overline{A}BC)} \tag{6.5}$$

The expansion in the denominator can be visualized by referring back to the $2 \times 2 \times 2$ joint probability table shown as Figure 5.2 and used in Exercise 5.9.12. The marginal sum, $P(BC)$, is the sum of cells 1 and 5, that is,

$$P(BC) = P(ABC) + P(\overline{A}BC)$$

Or, one could lean on the binary operations for logic,

$$P[\,((A \wedge B) \wedge C) \vee ((\overline{A} \wedge B) \wedge C)\,] = P[\,BC(A \vee \overline{A})\,] = P(BC) \times 1 = P(BC)$$

Contrary to the usual practice, we find that keeping to joint statements on the right hand side of Bayes's Theorem is more helpful than transforming them to conditional statements. The justification for this comes later when we show that the practical algorithm for numerical assignments[1] is designed for joint and marginal statements as opposed to conditional statements.

6.4   Generalizing the Number of Statements

So far, statements A, B, C, $\cdots$, were broken down into just two categories. For example, A and $\overline{A}$ are the two categories where statement A is TRUE and statement A is FALSE. And, correspondingly, we have shown only 2×2 joint probability tables for A and B, or $2 \times 2 \times 2$ joint probability tables for A, B, and C. However, we could let $\overline{A}$ itself be divided into two categories in a recursive manner.

As a change, suppose we want three categories for A. We might label these statements generically as $(A = a_1)$, $(A = a_2)$, and $(A = a_3)$. B is kept at just two categories. The corresponding 3×2 joint probability table would consist of three columns for A and two rows for B.

It is important to remember that we still maintain the requirement that each of these three statements in A can only be TRUE or FALSE. We wrap the probability operator around such statements when we do not possess enough information to ascertain whether the statement is, in fact, TRUE or FALSE.

For example, we might be connoisseurs of Elizabethan literature, and interested in making inferences about the authorship of Shakespeare's plays. If we restricted ourselves to two categories, we might frame the statements, $A \equiv$ "Shakespeare wrote the plays attributed to him.", and $\overline{A} \equiv$ "Shakespeare did not write the plays attributed to him."

However, if we wanted to expand the realm of discourse, we might let $\overline{A}$ be broken down into two more categories, $(A = a_2) \equiv$ "Marlowe wrote the plays attributed to Shakespeare.", and $(A = a_3) \equiv$ "de Vere wrote the plays attributed to Shakespeare." Thus, $\overline{(A = a_1)} \equiv$ "It is false that Shakespeare wrote the plays

---
[1]The practical algorithm is based on the *Maximum Entropy Principle* which is treated in detail in Volume II and *Information Geometry* described in Volume III.

attributed to him." is "Either Marlowe or de Vere, but not both, wrote the plays attributed to Shakespeare."

These statements are mutually exclusive and exhaustive. This means that the joint statement, "Marlowe wrote the plays attributed to Shakespeare and de Vere wrote the plays attributed to Shakespeare." must be false. That is the mutually exclusive part. The exhaustive part means that one of these possibilities must be true. There are no other possibilities for authorship (in our contrived world) other than Shakespeare, Marlowe, or de Vere.

Expressed in probability notation, the property of being mutually exclusive is reflected in the set of probability assignments of 0,

$$P\left[\,(A = a_1) \wedge (A = a_2)\,\right] \quad = \quad 0$$

$$P\left[\,(A = a_1) \wedge (A = a_3)\,\right] \quad = \quad 0$$

$$P\left[\,(A = a_2) \wedge (A = a_3)\,\right] \quad = \quad 0$$

$$P\left[\,(A = a_1) \wedge (A = a_2) \wedge (A = a_3)\,\right] \quad = \quad 0$$

Because of mutual exclusivity, after invoking the theorem on distributing P across the $\vee$ operator, we have,

$$P\left[\,(A = a_1) \vee (A = a_2) \vee (A = a_3)\,\right] = P(A = a_1) + P(A = a_2) + P(A = a_3)$$

This, in turn, leads to the exhaustive property,

$$P(A = a_1) + P(A = a_2) + P(A = a_3) = 1$$

These same constraints involving statements that are mutually exclusive and exhaustive apply to any set of *joint* statements as well, although the notation and their equivalent expressions in words quickly becomes unwieldy. Thus, we might write out that the particular joint statement involving the third category for A and the first category for B is mutually exclusive of the joint statement involving the first category for A and the second category for B,

$$P\left[\,(A = a_3) \wedge (B = b_1)) \vee ((A = a_1) \wedge (B = b_2))\,\right] =$$

$$P(A = a_3, B = b_1) + P(A = a_1, B = b_2)$$

It is best to simply think of this as the addition of the probabilities in cell 3 and cell 4 of a six cell joint probability table.

All this requires that we be specific about which statement we are referring to. For example, generalize the denominator in Bayes's Theorem to,

$$P(A = a_3 \,|\, B = b_1) \;=\; \frac{P(A = a_3, B = b_1)}{P(B = b_1)}$$

$$P(B = b_1) \;=\; \sum_{j=1}^{3} P(A = a_j, B = b_1)$$

$$=\; \sum_{j=1}^{3} P(B = b_1 \,|\, A = a_j) \; P(A = a_j)$$

$$P(A = a_3 \,|\, B = b_1) \;=\; \frac{P(B = b_1 \,|\, A = a_3) \; P(A = a_3)}{\sum_{j=1}^{3} P(B = b_1 \,|\, A = a_j) \; P(A = a_j)}$$

6.5   Using Bayes's Theorem to Generalize Logic

In the next Chapter, I will explain in some detail how probability theory generalizes Classical Logic. In succeeding Chapters, we will also show how to generalize Cellular Automata from the perspective of probability. And to do these things we will make heavy use of Bayes's Theorem.

As a prelude to the coming generalization of Classical Logic, take note of the fact that what we have labeled generically as abstract statements, ($e.g.$, $A, B, C, \ldots$), could just as easily refer to either a logic function or to the arguments of such a function. Therefore, there is no problem with treating logic functions and arguments as givens and placing them to the right of the conditioned upon symbol.

Here is an example of what we are talking about. Let C stand for the statement that the logic function $A \to B$ is operative. Then, we ask the question, "How is A updated if B is assumed to be true? And, by the way, I would also like the implication captured by statement C to be assumed true as well."

The answer comes in the form of Bayes's Theorem, where we write, quite in line with everything discussed so far,

$$P(A \,|\, BC) = \frac{P(ABC)}{P(BC)}$$

Now explicitly show the $\wedge$ operator and substitute the logic function for C,

$$P(A \,|\, BC) = \frac{P(\,[\,A \wedge [\,B \wedge [\,A \to B\,]\,]\,])}{P(\,[\,B \wedge [A \to B\,]\,])}$$

These logic operations within the probability operator are all well–defined and, when carried out, result in a symbolic answer provided by Bayes's Theorem.

A slightly more complex example now follows, but still easily understood in principle. Let D and E stand for two statements asserting that the following two logic functions, $A \to B$ and $B \to C$, are to be used as part of an inference.

Once again, both of these implications are placed to the right of the conditioned upon symbol. How uncertain are we about the argument C in the second implication given that we are assuming that the argument A in the first implication, and both logic functions are all true?

Writing out Bayes's Theorem as the answer to this question, we find that,

$$P(C\,|\,ADE) = \frac{P(CADE)}{P(ADE)}$$

Expanding both the numerator and the denominator and explicitly showing the $\wedge$ operator, we have,

$$P(C\,|\,ADE) = \frac{P(C \wedge [\,A \wedge [\,[\,A \to B\,] \wedge [\,B \to C\,]\,]\,])}{P(A \wedge [\,[\,A \to B\,] \wedge [\,B \to C\,]\,])}$$

Classical Logic tells us that C is TRUE given these assumptions, so this application of Bayes's Theorem had better return a value of $P(C\,|\,ADE) = 1$ if probability really does generalize Classical Logic. The next Chapter is a detailed discussion of how to use Bayes's Theorem in this manner.

An alternative way of saying that some logic function is to be employed in making the inferences is to substitute the DNF expansion of the function in question. Thus, we would place the DNF expansion of $f_\star(A, B, C, \cdots)$ to the right of the conditioned upon symbol instead of the actual function itself. For example, $AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$ could be used instead of writing $A \to B$. By definition, the DNF represents those variable settings where the functional assignment is T. This is the technique we will employ in the next Chapter when we begin to generalize the Classical Logic functions.

In these easy introductory examples, it was not too difficult to appreciate the fact that the logic functions and other information given to us were sufficient to make a *deduction*. If A is true and A implies B, this makes B true. If B is true and B implies C, this makes C true. A deductive chain like this means that something like C is known with certainty. And something known with certainty must have a probability of 1 or 0.

Ultimately, we will want to make quite clear that probability takes us beyond deduction and permits us to make *inferences*. In this case, the information processor does NOT possess enough information to make a deduction. Equivalently stated, the IP does not calculate probabilities of 1 or 0 for some statement.

Nonetheless, the formal rules of manipulation like Bayes's Theorem can still be applied. Consequently, statements will end up with probabilities *between* 1 and 0. Such a number is analogous to any clear–cut logical conclusion and indicates, in some sense, how close we are permitted to come to certainty by the inferential process.

6.6   Solved Exercises for Chapter Six

Exercise 6.6.1: Use the words *conditional*, *joint*, and *marginal* to rephrase Bayes's Theorem in words.

Solution to Exercise 6.6.1

A *conditional* probability is a *joint* probability divided by a *marginal* probability.

Exercise 6.6.2: What is Bayes's Theorem for two variables in terms of the cells in the 2 × 2 joint probability table?

Solution to Exercise 6.6.2

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{\text{Cell 1}}{\text{Cell 1} + \text{Cell 2}}$$

Exercise 6.6.3: Can $\overline{B}$ be used as the conditioning statement in Bayes's Theorem?

Solution to Exercise 6.6.3

Certainly. Bayes's Theorem now reads,

$$P(A \mid \overline{B}) = \frac{P(A\overline{B})}{P(\overline{B})} = \frac{\text{Cell 3}}{\text{Cell 3} + \text{Cell 4}}$$

Don't confuse the prohibition of placing a contradictory model to the right of the conditioned upon symbol with $\overline{B}$. It could very well happen that "B is FALSE" is a true statement and therefore perfectly acceptable to appear on the right side of the conditioned upon symbol. If we were to condition on the statement $A \wedge \overline{A}$, then we have problems. This contradiction (logic function $f_1(A, B)$ in Chapter Two) can never assume the value T and thus can never be placed to the right of the conditioned upon symbol.

Exercise 6.6.4: In the Shakespeare example, what is $(A = a_2) \oplus (A = a_3)$?

Solution to Exercise 6.6.4

The logic function $\oplus$ is called the "exclusive OR." The functional assignment is T if the first argument is T and the second F, or, if the first argument is F and the second T. In words, this corresponds to the statements, "Marlowe wrote the plays and de Vere did not, or, Marlowe did not write the plays and de Vere did." In turn, these statements are the same as, "Shakespeare did not write the plays."

Exercise 6.6.5: To exploit Bayes's Theorem within the context of the Shakespeare example, furnish some statement that stands for B. Then, fill in a 2 × 3 joint probability table with plausible numerical assignments that capture some model's joint relationship between A and B.

Solution to Exercise 6.6.5

Suppose that B is a statement based on characteristic use of the English language. Furthermore, imagine that experts in Elizabethan literature have studied a number of short phrases similar to "a sea of troubles" and "bare bodkin." Let B stand for the statement, "These kind of phrases appear in a soliloquy."

Next, construct a joint probability table shown below as Figure 6.1 mirroring a particular's model supposition about our three candidates, and their use of these kinds of characteristic phrases. Such a joint probability table might reflect a model containing a tentative working hypothesis that Shakespeare used these kinds of phrases in soliloquies relatively more often than either Marlowe or de Vere.

	$A=a_1$ Shakespeare	$A=a_2$ Marlowe	$A=a_3$ de Vere	
$B=b_1$ used phrases	0.50 ₁	0.02 ₂	0.07 ₃	0.59
$B=b_2$ did not use phrases	0.30 ₄	0.03 ₅	0.08 ₆	0.41
	0.80	0.05	0.15	1.00

Figure 6.1: *Joint probability table for Shakespeare's plays.*

Exercise 6.6.6: Given that it is true that the phrases "a sea of troubles" and "bare bodkin" appear within a soliloquy in *Hamlet*, a play attributed to Shakespeare, what is the updated state of knowledge that it was, in fact, Marlowe who wrote the plays?

Solution to Exercise 6.6.6

The answer is provided by Bayes's Theorem.

$$
\begin{aligned}
P(A = a_2 \,|\, B = b_1) &= \frac{P(A = a_2, B = b_1)}{\sum_{j=1}^{3} P(A = a_j, B = b_1)} \\[2ex]
&= \frac{0.02}{0.50 + 0.002 + 0.07} \\[2ex]
&= 0.0339
\end{aligned}
$$

The probability that Marlowe wrote the plays decreases from 0.05 to 0.0339 by this model's assumptions. It gets even closer to 0 than before; our state of knowledge is updated to greater certainty that Marlowe did not write the plays attributed to Shakespeare.

Similar calculations must of course also show a decreased probability for de Vere as well, accompanied by an increased probability for Shakespeare. In summary, this one fact represented by B serves to shift our state of knowledge more in favor of Shakespeare, and less in favor of his two rivals, that he did, in fact, write the plays attributed to him.

Exercise 6.6.7: So, do you now walk away from this exercise believing more in Shakespeare's claim to authorship?

Solution to Exercise 6.6.7

No, you cannot. The important lesson is this. What we showed in the previous exercise were the consequences upon *assuming a particular model were true*. The particular model in question is, in fact, the one instantiated by the joint probability table showing one legitimate numerical assignment to all six joint probabilities.

What is critical is this. What bearing do the observed facts have on this model, *and on all the other models that might also be proposed*? If the "data" were to overwhelmingly support this one model to the exclusion of all the other contenders, then, and only then, could we walk away with the change in the state of knowledge as derived above. Otherwise, we must take into account the data's support for all of the other models as well as their predictions about the change in the state of knowledge.

All of this will be made clearer in other applications involving the formal rules for probability symbol manipulation. The focus will then be directed at how states of knowledge about the various proposed models can change. This effort will be taken up in earnest quite soon.

Exercise 6.6.8: In the next several exercises we will check the common sense believability of Bayes's Theorem on examples that are initially very easy, but gradually become more difficult. First up, what is $P(A\,|\,A)$?

Solution to Exercise 6.6.8

Obviously, if A is true, then $P(A)$ must be 1. Bayes's Theorem for this case is,

$$P(A\,|\,A) = \frac{P(AA)}{P(A)} = \frac{P(A)}{P(A)} = 1$$

Exercise 6.6.9: What is $P(A \,|\, A \wedge B)$?

Solution to Exercise 6.6.9

If the logic function $A \wedge B$ is the model, then A must be true. Consequently, $P(A) = 1$. Bayes's Theorem reports back,

$$P(A \,|\, AB) = \frac{P(AAB)}{P(AB)} = \frac{P(AB)}{P(AB)} = 1$$

Exercise 6.6.10: What is $P(A \,|\, \overline{A})$?

Solution to Exercise 6.6.10

If A is false, then $P(A) = 0$. Bayes's Theorem gives us the same answer,

$$P(A \,|\, \overline{A}) = \frac{P(A\overline{A})}{P(\overline{A})} = \frac{P(F)}{P(\overline{A})} = \frac{0}{P(\overline{A})} = 0$$

Exercise 6.6.11: How is the previous result different than $P(A \,|\, A \wedge \overline{A})$?

Solution to Exercise 6.6.11

The very first logic function, that is, the `FALSE` logic function is specified as given. Applying Bayes's Theorem in a very mechanical way, just as we have been doing all along, we find that,

$$P(A \,|\, A \wedge \overline{A}) = \frac{P(AA \wedge A\overline{A})}{P(A \wedge \overline{A})} = \frac{0}{0} \equiv \text{undefined}$$

Exercise 6.6.12: What is $P(A \,|\, A \vee B)$?

Solution to Exercise 6.6.12

It is given that A or B or both are true by specifying the `OR` logic function to the right of the conditioned–upon symbol. We will examine each of these three cases after the Bayesian result is found.

$$P(A \,|\, A \vee B) = \frac{P(AA \vee AB)}{P(A \vee B)}$$

$$= \frac{P(A \vee AB)}{P(A) + P(B) - P(AB)}$$

$$P(A \vee AB) = P(A) + P(AB) - P(AAB)$$

$$= P(A)$$

In the end, we are left with,

$$P(A \mid A \vee B) = \frac{P(A)}{P(A) + P(B) - P(AB)}$$

Here, for the first time, we don't have an immediate answer. If we were to substitute the following numerical assignments from a possible joint probability table of the OR function, we would have,

$$P(A \mid A \vee B) = \frac{2/3}{2/3 + 2/3 - 1/3} = 2/3$$

As we will now demonstrate, A will be TRUE in two out of the three cases mentioned above and FALSE in the final third case.

First, A could be TRUE and B FALSE, leading to $P(A) = 1$. Bayes's Theorem yields,

$$\frac{P(A)}{P(A) + 0 - 0} = \frac{P(A)}{P(A)} = 1$$

As a second case, both A and B are TRUE, leading once again to $P(A) = 1$. Bayes's Theorem yields,

$$\frac{P(A)}{P(A) + 1 - 1} = \frac{P(A)}{P(A)} = 1$$

Again, the correct answer.

Finally, in the third case, A could be FALSE and B TRUE, so that $P(A) = 0$. Substituting 0 into the numerator of Bayes's Theorem yields the correct answer in this case,

$$P(A \mid A \vee B) = 0$$

Bayes's theorem returns the correct answer of 0 or 1 depending on A and B.

This argument isn't quite correct as it stands because we should have specified what was given as assumed TRUE in these three cases to the right of the conditioned upon symbol as well. We will show the correct formulation in a later exercise.

Exercise 6.6.13: Discuss $P(A \mid A \oplus B)$ in exactly the same way as the previous exercise.

Solution to Exercise 6.6.13

The logic function $A \oplus B$, in other words, the "exclusive OR" function is given on the right side of the conditioned upon symbol. As discussed in the Shakespeare example, either A is TRUE and B is FALSE, or A is FALSE and B is TRUE. Now in the first case if A is TRUE, then $P(A) = 1$. In the second case, if A is FALSE, then $P(A) = 0$. Does Bayes's Theorem cover these extreme cases?

Substitute the DNF expansion for $A \oplus B$,

$$P(A \mid A \oplus B) \quad = \quad \frac{P\left[\, A \wedge (\overline{A}B \vee A\overline{B})\,\right]}{P(\overline{A}B \vee A\overline{B})}$$

$$= \quad \frac{P(A\overline{A}B \vee AA\overline{B})}{P(\overline{A}B \vee A\overline{B})}$$

$$= \quad \frac{P(A\overline{B})}{P(A\overline{B}) + P(\overline{A}B)}$$

Supposing an assignment of $1/2$ to both $P(\overline{A}B)$ and $P(A\overline{B})$ with 0s, of course, for $P(AB)$ and $P(\overline{A}\,\overline{B})$,

$$P(A \mid A \oplus B) = \frac{1/2}{1/2 + 1/2} = 1/2$$

The IP (information processor) really doesn't know very much about the truth of A given just that this particular logic function, the XOR function, is operative. This makes sense because this function has a functional assignment of T in two cases. In the first case, the arguments are $A = T, B = F$, and in the second case, the arguments are $A = F, B = T$. This is a recapitulation of what the DNF expansion of the XOR told us.

Now, in the first case, if A is TRUE, then $P(A\overline{B})$ must be TRUE as well. Likewise, $P(\overline{A}B)$ must be FALSE. Substituting into the formula resulting from Bayes's theorem,

$$P(A \mid A \oplus B) = \frac{P(A\overline{B})}{P(A\overline{B}) + P(\overline{A}B)} = \frac{1}{1+0} = 1$$

By returning $P(A) = 1$, we see that Bayes's Theorem is correct in this case.

In the second case, if A is FALSE, then $P(A\overline{B})$ must be FALSE. Likewise, $P(\overline{A}B)$ must be TRUE. Substituting into the formula resulting from Bayes's Theorem,

$$P(A \mid A \oplus B) = \frac{P(A\overline{B})}{P(A\overline{B}) + P(\overline{A}B)} = \frac{0}{0+1} = 0$$

By returning $P(A) = 0$, Bayes's Theorem is validated once again.

Exercise 6.6.14: What is $P[\,B \mid f_8(A, B)\,]$?

Solution to Exercise 6.6.14

Here we have changed things up a bit by asking for the updated state of knowledge about B given the assumed truth of a logic function presented in the alternative functional notation used in Chapter Two. If we refresh our memories by referring back to that Chapter, we recall that this logic function is the EQUAL operator. This is the function that is prominently featured in proving logical equivalencies.

Thus, similarly to the previous exercises, we are asking for $P(B \,|\, A \leftrightarrow B)$. If we look back at Table 2.6, we see that this function can be TRUE in only two cases: 1) if both A and B are TRUE, and 2) if both A and B are FALSE. Indeed, this is exactly what a tautology or logical equivalence means. So, in the first case, B is TRUE and Bayes's Theorem had better return $P(B) = 1$. In the second case, B is FALSE and Bayes's Theorem had better return $P(B) = 0$.

Substituting the DNF expansion for $A \leftrightarrow B$, Bayes's Theorem works out to,

$$
\begin{aligned}
P(B \,|\, A \leftrightarrow B) &= \frac{P(B \wedge (AB \vee \overline{A}\,\overline{B}))}{P(AB \vee \overline{A}\,\overline{B})} \\[2mm]
&= \frac{P(BAB \vee B\overline{A}\,\overline{B}))}{P(AB \vee \overline{A}\,\overline{B})} \\[2mm]
&= \frac{P(AB)}{P(AB) + P(\overline{A}\,\overline{B})}
\end{aligned}
$$

The IP finds itself in the same situation as with the XOR logic function. The IP's state of knowledge about the truth of B when the intuitive numerical assignments of $1/2$ are made for $P(AB)$ and $P(\overline{A}\,\overline{B})$ remains at $P(B \,|\, A \leftrightarrow B) = 1/2$. The EQUAL function has a functional assignment of T in two cases. In the first case, the arguments are $A = T, B = T$, and in the second case, the arguments are $A = F, B = F$. Here we see that B takes on the value T in one case and the value F in the other case. So it makes intuitive sense that $P(B \,|\, A \leftrightarrow B) = 1/2$.

If A and B are both TRUE, then $P(AB) = 1$ and $P(\overline{A}\,\overline{B}) = 0$. Substituting these values into Bayes's Theorem,

$$
P(B \,|\, A \leftrightarrow B) = \frac{P(AB)}{P(AB) + P(\overline{A}\,\overline{B})} = \frac{1}{1+0} = 1
$$

Thus, as we have come to expect by now, Bayes's Theorem does report back the correct answer.

Checking the second case where both A and B are FALSE, we bump into the symmetrical outcome. $P(AB) = 0$ and $P(\overline{A}\,\overline{B}) = 1$. Substituting these values into Bayes's Theorem,

$$
P(B \,|\, A \leftrightarrow B) = \frac{P(AB)}{P(AB) + P(\overline{A}\,\overline{B})} = \frac{0}{0+1} = 0
$$

So Bayes's Theorem returns $P(B) = 1$ or $P(B) = 0$ in each case as the DNF expansion requires.

Exercise 6.6.15: Here is the last exercise of this type. What is the probability that B is FALSE if we are assuming that the NOR logic function is specified as the model for the numerical assignments?

Solution to Exercise 6.6.15

The definition of NOR really answers the question. This logic function can only have the assignment of T in one case and that unique situation demands that the variable assignment to both A and B must be F. Refer back to Table 2.6 where the DNF for $f_2(A, B)$ was $\overline{A}\,\overline{B}$. Therefore, if the NOR logic function is in charge (that is, if it is given as the model making the numerical assignments), B must be FALSE and $P(B) = 0 \equiv P(\overline{B}) = 1$. We fully expect that Bayes's Theorem will provide the same answer.

$$
\begin{aligned}
P(\overline{B}\,|\,A \downarrow B) &= \frac{P(\overline{B} \wedge (A \downarrow B))}{P(A \downarrow B)} \\[2mm]
&= \frac{P(\overline{B} \wedge (\overline{A}\,\overline{B}))}{P(\overline{A}\,\overline{B})} \\[2mm]
&= \frac{P(\overline{B}\,\overline{A}\,\overline{B}))}{P(\overline{A}\,\overline{B})} \\[2mm]
&= \frac{P(\overline{A}\,\overline{B})}{P(\overline{A}\,\overline{B})} \\[2mm]
&= 1
\end{aligned}
$$

By the **Sum Rule**,

$$
\begin{aligned}
P(B\,|\,A \downarrow B)) &= 1 - P(\overline{B}\,|\,A \downarrow B) \\[2mm]
&= 0
\end{aligned}
$$

Here we see an illustration of the fact that the formal manipulation rules require that $P(B\,|\,\star) + P(\overline{B}\,|\,\star) = 1$ no matter what numerical value might be assigned to B because of the conditioning information $\star$ appearing to the right of the conditioned upon symbol. All of the formal rules for probability symbol manipulation that we have developed work irrespective of the fact that the conditioning information might change. The consequence is that the abstract probabilities might possess different numerical assignments, but the **Sum Rule** remains inviolable.

Exercise 6.6.16: Assess the intuitive reasonableness of Bayes's Theorem under these conditions. Suppose that there are only two models under consideration with nothing to distinguish whether one is better than the other. In an attempt to rectify this situation, data are gathered, but, unfortunately, these data do not discriminate between the two models. Cast this scenario into expressions amenable to treatment by Bayes's Theorem.

Solution to Exercise 6.6.16

An initial state of knowledge concerning the two models when the IP is completely uninformed about the relative status of the two models is captured by,

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = 1/2$$

Now suppose, for the sake of illustration, that the probability of the data given that model 1 is actually true is $P(\mathcal{D} \,|\, \mathcal{M}_1) = 0.04$.

In the statement of the problem, we specified that the data do not discriminate between the two models. This means that the probability of the data given that the competing model, model 2, is actually true is also $P(\mathcal{D} \,|\, \mathcal{M}_2) = 0.04$.

The question we now ask is, "What is the updated state of knowledge about these two models?" Applying Bayes's Theorem, we have,

$$
\begin{aligned}
P(\mathcal{M}_1 \,|\, \mathcal{D}) \;&=\; \frac{P(\mathcal{M}_1, \mathcal{D})}{P(\mathcal{D})} \\[2mm]
&=\; \frac{P(\mathcal{D} \,|\, \mathcal{M}_1)\, P(\mathcal{M}_1)}{P(\mathcal{D})} \\[2mm]
&=\; \frac{P(\mathcal{D} \,|\, \mathcal{M}_1)\, P(\mathcal{M}_1)}{P(\mathcal{D} \,|\, \mathcal{M}_1)\, P(\mathcal{M}_1) + P(\mathcal{D} \,|\, \mathcal{M}_2)\, P(\mathcal{M}_2)} \\[2mm]
&=\; \frac{0.04 \times 1/2}{(0.04 \times 1/2) + (0.04 \times 1/2)} \\[2mm]
&=\; 1/2
\end{aligned}
$$

The application of Bayes's Theorem has returned the eminently intuitive result that the IP's state of knowledge about the models has not changed one whit from what it was before the data were observed. We also have another example of the fact that the **Sum Rule** works irrespective of the conditioning statements,

$$P(\mathcal{M}_1 \,|\, \mathcal{D}) + P(\mathcal{M}_2 \,|\, \mathcal{D}) = P(\mathcal{M}_1) + P(\mathcal{M}_2) = 1$$

This exercise foreshadows extensive discussions of two foundational questions: 1) what is a "prior probability?", and 2) what does an "uninformed" IP mean?

Exercise 6.6.17: Seemingly every Bayesian textbook presents a variant of the following scenario: You are tested for a rare disease by a clinical test which is 95% "correct." The false positive rate of the test, that is, the proportion of times it returns a positive indication when, in fact, the disease is not present, is 2%. What is your updated state of knowledge that you have the disease given that your test came back positive?

Solution to Exercise 6.6.17

The solution to this problem follows the same Bayes's Theorem template we have been relying on for all of the previous exercises. At the outset, before plugging in any numbers, we want to belabor once more a persistent bugaboo. It is a fundamentally simple conceptual notion that seems to cause an unwarranted degree of confusion.

Here is that fundamental concept. Bayes's Theorem is a formal manipulation rule for abstract probability symbols. As such, it is valid for all legitimate numerical values assigned to these probability symbols.

Thus, it might happen that legitimate values are assigned under some model that you happen to disagree with. But that is not the concern of Bayes's Theorem. It will process any and all legitimate numerical assignments correctly. It is **NOT** incumbent upon Bayes's Theorem to remark on the sense or nonsense of your or my particular numerical assignments.

It is important then not to confuse any potentially deserved criticism of the numerical assignments with an improper indictment of Bayes's Theorem as a formal manipulation rule. One must be careful to disentangle what a *Bayesian analysis* really means.

For me, it means that Bayes's Theorem has been applied properly to achieve the goal of revising probabilities based on what statements are taken to be true, and not that the numerical assignment under some model **MUST** be correct.

Having released my spleen with this diatribe, let us now plug in the numbers as given in the statement of the problem. After the result, we'll continue the discursive discussion a bit longer. Look at Figure 6.2 at the top of the next page which shows a *contingency table.*

A contingency table displays actual *frequencies* in each cell as opposed to a joint probability table which shows *probabilities* for joint statements in the cells. Suppose, for the sake of the example, that 10,000 people have been tested for the disease. The table shows the observed frequencies for the four possible combinations of disease present or absent and test positive or negative together with the marginal frequencies.

It is important to constantly keep in mind that frequencies are not probabilities. On the other hand, there is nothing to prevent some model from assigning numerical values to probabilities that match normed frequency counts. In a rather peculiar way, we want to emphasize the difference between these two concepts by assigning numerical values to the joint probabilities that match the normed frequencies that appear in the contingency table.

Figure 6.2: *Contingency table showing the frequencies for a test of a disease.*

I do this with great trepidation. It is absolutely imperative that this particular numerical assignment be understood as contingent on assuming some particular k^{th} model $\mathcal{M}_k$ to be true. When we do this, we are, in effect, ignoring an infinite number of other models that would be making other numerical assignments.

With that caveat ringing in our ears, insert the following probability assignments into Bayes's Theorem to obtain a revised probability of having the disease given a positive result on the test,

$$
\begin{aligned}
P(A \mid B, \mathcal{M}_k) &= \frac{P(AB \mid \mathcal{M}_k)}{P(B \mid \mathcal{M}_k)} \\[2mm]
&= \frac{0.0095}{0.0293} \\[2mm]
&= 0.3242
\end{aligned}
$$

If you have a positive result from the test, your state of knowledge about having the disease is revised considerably upwards from the original general prevalence of the disease which exists at 1%. However, it is not quite as bad as you first thought given that 95% hit rate.

That's due to that seemingly small error rate of 2% for the test (that is, 198 patients divided by 9900) for whom a positive result was reported even when they didn't have the disease . There are so many more people who don't have the disease that you might take solace in the fact that you might be like one of the 198 people in the $\overline{A}B$ cell of the contingency table.

The joint probability was used in the numerator Bayes's Theorem above. The alternative version uses conditional probabilities. For clarity, conditioning on $\mathcal{M}_k$ is suppressed. Of course, the two ways had better return the same answer.

$$
P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B \mid A)\,P(A) + P(B \mid \overline{A})\,P(\overline{A})}
$$

$$P(A \mid B) = \frac{0.95 \times 0.01}{(0.95 \times 0.01) + (0.02 \times 0.99)}$$

$$= 0.3242$$

The numerator, $P(AB)$, is converted to $P(B \mid A)\, P(A)$ by **Commutativity** and the **Product Rule**. The denominator, $P(B)$, is seen to be the marginal sum of $P(AB)$ and $P(\overline{A}B)$. When these two terms are each subjected to **Commutativity** and the **Product Rule** we have,

$$P(B) = P(B \mid A)\, P(A) + P(B \mid \overline{A})\, P(\overline{A})$$

While we are deep into this problem, here is an interesting question. Ask yourself what initial assignment of uncertainty to the prevalence of the disease would actually result in an updated probability of having the disease close to 0.95. It turns out that, if the assignments $P(A) = 1/3$ and $P(\overline{A}) = 2/3$ are made, then applying Bayes's Theorem in the same manner as just shown would revise the probability of your having the disease contingent upon a positive test result to about 0.96.

Now, as alluded to at the outset, you may not like this outcome. But Bayes's Theorem, and its use within this kind of circumscribed Bayesian analysis, is immune to any such criticism! What you don't like is something else. If you think that the updated state of knowledge concerning your having the disease is too high, then it is my assignment of numerical values under some model $\mathcal{M}_k$ that you are to find fault with.

Most likely, you would not quibble with the assignments to $P(A \mid B)$ and $P(A \mid \overline{B})$ but rather with the initial assignments to $P(A)$ and $P(\overline{A})$. And, in fact, every text-book presentation makes some kind of numerical assignment closely allied to the *frequency* counts as I did above with Figure 6.2.

Suppose that no frequency table of this kind has been provided. I might try to justify my larger numerical assignment to $P(A)$ and $P(\overline{A})$ in the following manner. I assume that no information of any kind has been provided about you.

For a disease like AIDS certain relevant factors are known to play some kind of role. Presumably, I don't know if you are male or female, black or white, old or young, an IV drug abuser or not, a celibate monk or habitué of gay bars. If I knew any or all of these relevant factors, I would adjust my numerical assignment appropriately. But lacking any of this knowledge, it is more of a coin toss if you have the disease.

Conversely, by what right does anyone assign an extremely low value to your having the disease based on a presumed *lack of knowledge*? It seems to me that in order to assign such a low value I would need to know quite a bit about you. And if I were to find out that you are 90 year old celibate monk that has lived on a deserted island for the past 60 years, I would have no problem assigning an extremely low numerical value to the abstract probability represented by $P(A)$.

But, in the end, we see that none of us is quibbling over the Bayesian analysis. We just have defensible disagreements on the numerical assignments, which Bayes's Theorem gives us every leeway to pursue.

Exercise 6.6.18: In the next Chapter, the need will arise to have a version of Bayes's Theorem for three variables that keeps the third variable to the right of the conditioned upon symbol for all expressions. Derive this alternative version of Bayes's Theorem in the same manner as was shown for two variables.

Solution to Exercise 6.6.18

The following version of Bayes's Theorem is employed in the next Chapter in order to find $P(B \mid AC)$ where C is to be kept as a given to the right of the conditioned upon symbol.

$$P(B \mid AC) = \frac{P(AB \mid C)}{P(A \mid C)}$$

With only two variables A and B, there were only two permutations possible for the **Commutativity axiom** to operate on. Thus, $P(AB) = P(BA)$ could start the derivation of Equation (6.3). With three variables, A,B, and C, there are $3! = 6$ permutations, (1) ABC, (2) ACB, (3) BAC, (4) BCA, (5) CAB, and (6) CBA.

Through **Commutativity** and **Associativity**, the joint probability for all six permutations are equal. It is straightforward, but tedious, to continuously apply these two axioms at a number of steps to finally rearrange the variables into the desired order. Here is a condensed example.

$$
\begin{aligned}
C(BA) &= C(AB) \\
&= (CA)B \\
&= (AC)B \\
&= A(CB) \\
&= A(BC)
\end{aligned}
$$

Pick out the following two equal joint probabilities so that a convenient cancelation can occur when the **Product Rule** is applied.

$$P(BAC) = P(ABC)$$

Now apply the **Product Rule** to both sides of the equation,

$$P(B \mid AC)\, P(A \mid C)\, P(C) = P(A \mid BC)\, P(B \mid C)\, P(C)$$

Divide both sides by $P(C)$ for that aforementioned desired cancelation,

$$P(B \mid AC) \, P(A \mid C) = P(A \mid BC) \, P(B \mid C)$$

All that's left is to do is divide through by $P(A \mid C)$, in order to turn the right hand side into the expression we were looking for in Bayes's Theorem,

$$P(B \mid AC) = \frac{P(A \mid BC) \, P(B \mid C)}{P(A \mid C)}$$

The **Product Rule** can be used in reverse when the conditional probabilities are used in the numerator to arrive at the version in which it appears in the next Chapter.

$$P(B \mid AC) = \frac{P(AB \mid C)}{P(A \mid C)}$$

As a check on the internal consistency of these manipulations, substitute the right hand side of Bayes's Theorem for all the conditional probabilities appearing in both the numerator and denominator on the right hand side of the above equation,

$$P(B \mid AC) = \frac{P(A \mid BC) \, P(B \mid C)}{P(A \mid C)}$$

Carry out all the cancelations, and then apply **Commutativity** and **Associativity** one last time,

$$
\begin{aligned}
P(B|AC) &= \frac{\frac{P(ABC)}{P(BC)} \times \frac{P(BC)}{P(C)}}{\frac{P(AC)}{P(C)}} \\[2em]
&= \frac{P(ABC)}{P(AC)} \\[2em]
&= \frac{P(BAC)}{P(AC)}
\end{aligned}
$$

Chapter 7

Generalizing Logic with Probability

7.1 Introduction

The developments in this Chapter are the payoff for the long, slow buildup of the supporting ideas that we have been discussing. We are able to demonstrate that the formal rules for manipulating probability symbols generalize Classical Logic. This is the foundation for bridging the gap between logical deduction and probabilistic inference.

Jaynes [12] gave a few examples of how Classical Logic can be generalized by probability theory. Any generalization must reproduce known results from the subject which it claims to generalize. Thus, probability theory must first be able to reproduce results from Classical Logic.

Jaynes illustrates how probability theory correctly generalizes what he calls the "strong syllogisms" taken from Classical Logic. He then moves on to a discussion of "weak syllogisms" which Classical Logic calls "invalid." This kind of classically invalid reasoning, if viewed more generously within the context of probability theory, can be understood not as "invalid," but as a case of missing information.

A slightly more expanded explanation is put on offer to justify probability as a generalization of logic. We do so by first building on the purely formal manipulation rules that have been developed. As an additional aid to understanding, we also make heavy use of joint probability tables with legitimate numerical assignments inserted into the cells of the table.

At the end of the Chapter, I indulge a pet peeve of mine. The reasoning employed in criminal investigations and judicial pronouncements has always seemed a bit suspect to me. Consequently, I broach the topic of *circumstantial evidence* to see what inference as a generalization of logic has to say about this topic.

7.2   Preparing for the Generalization

There is a little bit of groundwork to do before we plunge into the task of generalizing Classical Logic. By now, we are comfortable with the circumscribed universe of Classical Logic as represented by the sixteen Boolean functions of two variables.

In preparation, we might think in terms of *three* statements, labeled as A, B, and Z. A and B remain as the usual notation for the two arguments of a logic function, while Z is a statement concerning the functional assignment made by some particular logic function.

There are four possibilities for the arguments A and B. We choose to list them conventionally in the order (1) TT, (2) TF, (3) FT, and (4) FF. The `CONDITIONAL` operator, as one example of a logic function from the total of sixteen available, was defined by the following functional assignments to each of these four variable settings.

1. $f_{13}(T,T) = T$,

2. $f_{13}(T,F) = F$,

3. $f_{13}(F,T) = T$,

4. $f_{13}(F,F) = T$

So, $Z = T$ for the variable settings in (1), (3), and (4), and $Z = F$ for the variable settings in (2). This functional assignment led directly to the full DNF expansion, or, if you prefer, the orthonormal function expansion of the implication function as,

$$f_{13}(A,B) \equiv A \to B \equiv AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$$

Notationally, the implication function $A \to B$ was introduced as $f_{13}(A,B)$ in Chapter Two with the label of the `CONDITIONAL` operator. In *Mathematica* syntax, this logic function is expressed as `Implies[p, q]` where the arguments `p` and `q` can take on the values of `True` or `False`. The function `Implies[p, q]` will then be evaluated and return with either `True` or `False`.

When proving probability theorems about Classical Logic, we want the freedom to write out Bayes's Theorem with different looking, but equivalent, expressions appearing to the right of the conditioned upon symbol. In one case, we might write the logic function itself as in $P(B \,|\, A, A \to B)$.

Boole's Expansion Theorem permits us write the implication function as,

$$
\begin{aligned}
f_{13}(A,B) &= f(T,T)\,AB \vee f(T,F)\,A\overline{B} \vee f(F,T)\,\overline{A}B \vee f(F,F)\,\overline{A}\,\overline{B} \\[2mm]
&= (T \wedge AB) \vee (F \wedge A\overline{B}) \vee (T \wedge \overline{A}B) \vee (T \wedge \overline{A}\,\overline{B}) \\[2mm]
&\equiv AB \vee \overline{A}B \vee \overline{A}\,\overline{B}
\end{aligned}
$$

Thus, as a second way of expressing things, substitute the DNF expansion for the logic function,

$$P(B \mid A, AB \vee \overline{A}B \vee \overline{A}\,\overline{B})$$

Or, in a third case, we might prefer a more general version using $\mathcal{M}_k$ to stand for the numerical assignment consistent with the logic function, as in,

$$P(B \mid A, \mathcal{M}_k)$$

The notation $\mathcal{M}_k$ indicates that some k^{th} model has assigned numerical values to the joint probability table. These numerical values will lead to the same answers as the logic function.

Finally, to simplify the notation even further,[1] we will often simply write,

$$P(B \mid A, Z)$$

By definition, if $Z = T$, the DNF expansion and $Z = T$ mean the same thing, so we can abbreviate longer expressions to,

$$P(B \mid A, AB \vee \overline{A}B \vee \overline{A}\,\overline{B}) \equiv P(B \mid A, Z)$$

with the idea that we can substitute the proper DNF expansion wherever Z appears.

The rest of the Chapter presents the detailed arguments for probability as a generalization of Classical Logic. Logic has been a central philosophical concern for the human mind since antiquity. Aristotle has been credited with the first systematic codification; as evidence, the extensive use of the word *syllogism* in logic. When medieval theologians were not debating the number of angels who could dance on the head of a pin, they were giving names to some of the well known syllogisms of logic. Since Latin was the *lingua franca* for the educated class at that time, we recognize why classical Latin labels became attached to some of the following examples.

7.3 Strong Syllogisms

7.3.1 *Modus ponens*

For the first example of a strong syllogism, consider the classical syllogism, otherwise known as *modus ponens*,[2] $(A \wedge [\ A \to B\ ]) \to B$. In words, if A is true, and A implies B, then both of these together imply that B is true as well. Let another statement, labeled as Z, be the statement that $A \to B$ is the particular logic function assumed as true.

---

[1] Pun intended.

[2] Latin for *A way of affirming*. Jaynes was strongly influenced by Pólya's book *Patterns of Plausible Inference* and so is the treatment in this Chapter.

In the notation of probability theory, we would write the classical syllogism just given as $P(B \mid A, Z)$, the probability that B is TRUE given that we assume A is TRUE. $Z = T$ indicates that it is permissible to substitute the DNF expansion for the particular logic function. $P(B \mid A, Z)$ must then equal 1 to reproduce the result from Classical Logic.

From Bayes's Theorem, we know that the probability of B, assuming that A as well as the logic function indicated by Z are both TRUE, is given by,

$$P(B \mid A, Z) = \frac{P(BAZ)}{P(AZ)}$$

The implication $A \rightarrow B$ is, of course, one of the 16 logic functions involving two variables. In our alternative notation, the implication was listed as $f_{13}(A, B)$. Earlier, we demonstrated the very important fact that any one of these functions could be expressed in the disjunctive normal form. From Table 2.6, the DNF for the implication, $f_{13}(A, B)$, is,

$$(Z = T) \equiv A \rightarrow B \equiv (A \wedge B) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$$

When we condense the DNF expansion by eliminating the explicit reference to the $\wedge$ operator, we write,

$$A \rightarrow B \equiv AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$$

This full DNF expansion for the $\rightarrow$ operator may be expressed in a shorter way. Using this shorter version will make the proofs easier. The next section presents a derivation utilizing this new expression for implication. We will immediately take advantage of this effort by substituting it for the full DNF expansion for implication.

7.3.2   A shorter formula for implication

We have in our possession the main result from Bayes's Theorem where Z appears in a joint statement with A and B. But that in itself is not particularly enlightening. As a first step towards winding out the full import of this rather compact macro, we show the logical equivalence between the full DNF expression and a shorter expression for the implication binary operator.

The full DNF for logical implication is again,

$$Z \equiv AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$$

Focus now on the last two terms where the common factor of $\overline{A}$ can be extracted,

$$\overline{A}B \vee \overline{A}\,\overline{B} = \overline{A} \wedge (B \vee \overline{B})$$

and then, because $B \vee \overline{B} = T$, this result further reduces the last two terms to,

$$\overline{A}B \vee \overline{A}\,\overline{B} = \overline{A}$$

Place $\overline{A}$ in front of the first term, and use the **Distributivity axiom** to form,

$$\overline{A} \vee (A \wedge B) = (\overline{A} \vee A) \wedge (\overline{A} \vee B)$$

Next, use the **Complementation axiom** to change the first term to T and arrive at the desired result of,

$$T \wedge (\overline{A} \vee B) = \overline{A} \vee B \equiv AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$$

7.3.3 Continuing with the proof

By making this substitution for $A \rightarrow B$ into Bayes's Theorem, we have managed to unwind the opaque general form to,

$$P(B \,|\, A, A \rightarrow B) \equiv P(B \,|\, A, \overline{A} \vee B)$$

We are now properly positioned to carry out the Boolean operations indicated by Bayes's Theorem in both the numerator and denominator.

$$P(B \,|\, A, Z) = \frac{P(B \wedge [\, A \wedge [\, \overline{A} \vee B \,]\,]\,)}{P(A \wedge [\, \overline{A} \vee B \,]\,)}$$

Concentrating first on the easier denominator,

$$P(A \wedge [\, \overline{A} \vee B \,]\,) \;\; = \;\; P(A\overline{A} \vee AB)$$

$$= \;\; P(AB)$$

With this result in hand, the numerator is,

$$P(B \wedge [\, A \wedge [\, \overline{A} \vee B \,]\,]\,) \;\; = \;\; P(B \wedge [\, AB \,]\,)$$

$$= \;\; P(AB)$$

Substitute these findings back in for the numerator and denominator of Bayes's Theorem to see that,

$$P(B \,|\, A, Z) = \frac{P(AB)}{P(AB)} = 1$$

We have achieved our goal of showing that B must be TRUE given that A is TRUE. Not to be forgotten in all of this is the additional important fact that the implication operator is the deductive model being used. Probability symbol manipulation dictates that B must be TRUE if we assume that A is TRUE and Z is TRUE. Probability theory as a generalization must return, and does return, the same answer as Classical Logic.

7.3.4   A joint probability table for *modus ponens*

To augment this purely formal approach, it always helps to have another way of
seeing how the result is arrived at. To that end, Figure 7.1 below shows a conceivable
joint probability table for the three statements A, B, and Z. There are some
representative numbers inserted into each of the eight cells of the table. Each of
these numbers is a legitimate probability assignment to the joint occurrence indexed
by that cell.

Figure 7.1: $2 \times 2 \times 2$ *joint probability table to illustrate how probability theory can
reproduce logical implication.*

The statement Z represents the case when the functional assignment is for the
specified logic function f_{13}, while the statement $\overline{Z}$, "Z is FALSE," represents the
functional assignment from f_4. Each of the eight cells indexes a joint statement
where, for example, cell 1, ZAB, is the statement, "$Z = T$ and $A = T$ and $B = T$."
and cell 8, $\overline{Z}\,\overline{A}\,\overline{B}$, is the statement, "$Z = F$ and $A = F$ and $B = F$." The numbers
placed into each cell are legitimate probabilities under Z and $\overline{Z}$.

However, given that implication is the logic function being used, Z cannot be
TRUE if A is TRUE and B is FALSE. That would violate the very definition of
what implication means from the standpoint of Classical Logic. Cell 3 indexes this
particular setting for the three variables where an F must be assigned. Cell 3,
therefore, will have a numerical assignment of $P(ZA\overline{B}) = 0$.

The statement $\overline{Z}$ is, "The implication logic function $f_{13}(A, B)$ is not the model
assigning numerical values to the joint statements." The mirror symmetry reflected
in choosing the logic function $f_4(A, B)$ for $\overline{Z}$ means that the DNF for this function
has coefficients just the opposite of $f_{13}(A, B)$. Zeroes are placed into those cells
where positive values appeared under f_{13} and *vice versa*.

We have set up this joint probability table to express a state of knowledge about any joint statement involving three statements. Additionally, each statement could assume only two values. The sum of the numerical assignments placed into all eight cells must equal 1. That is, one of these eight joint statements must happen. Figure 7.1 also shows the marginal probabilities for each of the three variables.

$$P(A) = P(\overline{A}) = P(B) = P(\overline{B}) = 1/2$$

and $P(Z) = 3/4$ with $P(\overline{Z}) = 1/4$.

We might intuitively justify these particular numerical assignments through an appeal to symmetry. A and B both take on T and F equally often. However, invoking the implication logic function means that T is going to be the value of the functional assignment Z for three out of the four possibilities that the arguments can assume.

In addition, there are marginal probabilities that reflect the joint probabilities of the two arguments A and B as summed over Z. For example,

$$P(AB) = P(ZAB) + P(\overline{Z}AB) = 1/4 + 0 = 1/4$$

Thus, $P(AB)$ is the marginal sum of cells 1 and 5.

It is easy to numerically verify the results found through formal manipulation of the symbols. In the previous section, we just proved that,

$$P(B \,|\, A, A \rightarrow B) = \frac{P(AB)}{P(AB)} = 1$$

Working directly from,

$$P(B \,|\, A, Z) = \frac{P(ZAB)}{P(ZA)} = \frac{P(ZAB)}{P(ZAB) + P(ZA\overline{B})}$$

substitute the numerical assignment appearing in cell 1 for the joint probability in the numerator, and then the sum of the numerical assignments appearing in cells 1 and 3 for the marginal probability of the denominator,

$$P(B \,|\, A, A \rightarrow B) = P(B \,|\, A, Z) = \frac{P(ZAB)}{P(ZAB) + P(ZA\overline{B})} = \frac{1/4}{1/4 + 0} = 1$$

7.3.5 *Modus tollens*

We will go ahead and show a second example of a strong syllogism, *modus tollens*,[3] which probability theory also correctly generalizes. Written in binary operator notation, this strong syllogism is expressed as,

$$(\overline{B} \wedge [\, A \rightarrow B \,]) \rightarrow \overline{A}$$

---

[3]Latin for *A way of denying*. Again, discussed as a precursor to inferential reasoning by Pólya in *Patterns of Plausible Inference*.

In words, if B is FALSE and A implies B, then both of these together imply that A is also FALSE. The statement Z remains the same as before, that is, $A \rightarrow B$.

In the notation of probability theory, we would write this second strong classical syllogism as $P(A \,|\, \overline{B}, Z)$, the probability that A is true given that we assume that B is FALSE and Z is TRUE, or, equivalently, that we are substituting the DNF expansion for this function. Therefore, $P(A \,|\, \overline{B}, Z)$ must equal 0, or $P(\overline{A} \,|\, \overline{B}, Z)$ must equal 1, to reproduce the result from Classical Logic.

We'll follow the same presentation order by showing the formal manipulation rules first and the joint probability table solution second. The formal manipulation rules give us,

$$P(A \,|\, \overline{B}, Z) = \frac{P(A \wedge [\, \overline{B} \wedge [\, \overline{A} \vee B\,]\,])}{P(\overline{B} \wedge [\, \overline{A} \vee B\,])}$$

Focus first on the denominator. Apply the **Distributivity axiom** followed by the **Complementation** and **Commutativity axioms** to arrive at the probability for the denominator as,

$$P(\overline{B} \wedge [\, \overline{A} \vee B\,]) \;\; = \;\; P(\overline{B}\,\overline{A} \vee B\overline{B})$$

$$= \;\; P(\overline{A}\,\overline{B})$$

Next, carry out the $A \wedge [\, \cdots \,]$ operation in the numerator on the result just found,

$$P(A\overline{A}\,\overline{B}) = P(F)$$

Bayes's Theorem now looks like,

$$P(A \,|\, \overline{B}, Z) = \frac{P(F)}{P(\overline{A}\,\overline{B})} = 0$$

since $P(F)$ is represented by 0.

Things have ended up satisfactorily. The manipulation rules for probability have verified the result that it is impossible for A to be TRUE if it is given that B is FALSE and the implication logic function is the deductive model.

7.3.6 Joint probability table solution for *modus tollens*

The same joint probability table used to illustrate the *modus ponens* example can do double duty for *modus tollens*. Z is the statement that is conditioned upon, that is, the statement we are (subjunctively) assuming as true to see what its consequences might be. And, of course, Z is still the particular logic function called implication. Consequently, cell 3 must contain a numerical assignment of 0 if Z is assumed true.

Directly invoking Bayes's Theorem in order to express the right hand side in terms of the joint and marginal probabilities,

$$P(A \,|\, \overline{B}, Z) = \frac{P(ZA\overline{B})}{P(Z\overline{B})}$$

The denominator is expanded to,

$$P(Z\overline{B}) = P(ZA\overline{B}) + P(Z\overline{A}\,\overline{B})$$

Now substitute the numerical assignments in cells 3 and 4 from Figure 7.1,

$$P(A \,|\, \overline{B}, Z) = \frac{0}{0 + 1/4} = 0$$

This alternative solution confirms the answer we obtained above in the purely formal approach.

Notice that in the formal approach, the denominator was $P(\overline{A}\,\overline{B})$ while in the solution using the joint probability table the denominator was $P(Z\overline{B})$. But we can plug in the numerical values to see that in the first instance,

$$P(\overline{A}\,\overline{B}) = P(Z\overline{A}\,\overline{B}) + P(\overline{Z}\,\overline{A}\,\overline{B}) = 1/4 + 0 = 1/4$$

while in the second instance,

$$P(Z\overline{B}) = P(ZA\overline{B}) + P(Z\overline{A}\,\overline{B}) = 0 + 1/4 = 1/4$$

7.3.7 Checking the Product Rule

It is a good idea to occasionally take stock and check that the formal rules are not being violated in what we have done so far. As an example, a calculation should show that the joint probability $P(ZAB)$ is equal to its decomposition via the **Product Rule**. After invoking the **Commutativity axiom** twice, and the **Associativity** axiom once, the **Product Rule** decomposition for $P(ZAB)$ is,

$$P(ZAB) = P(ABZ) = P(A \,|\, B, Z)\, P(B \,|\, Z)\, P(Z)$$

Using the numerical assignments in the joint probability table of Figure 7.1, start working from the rightmost term where the marginal sum for $P(Z)$ can be read directly as 3/4. For the next term, use Bayes's Theorem to find

$$P(B \,|\, Z) = \frac{P(BZ)}{P(Z)} = \frac{1/4 + 1/4}{3/4} = 2/3$$

The third and final term is also found easily by Bayes's Theorem as,

$$P(A \,|\, B, Z) = \frac{P(ZAB)}{P(BZ)} = \frac{1/4}{1/4 + 1/4} = 1/2$$

The final multiplication of these three terms must equal the numerical assignment for the probability in cell 1, $P(ZAB)$,

$$P(ZAB) = 1/2 \times 2/3 \times 3/4 = 1/4$$

Of course, this is just a reflection of the cancelations involved in,

$$
\begin{aligned}
P(A, B, Z) &= P(A \,|\, B, Z)\, P(B \,|\, Z)\, P(Z) \\[2mm]
&= \frac{P(ABZ)}{P(BZ)} \times \frac{P(BZ)}{P(Z)} \times P(Z)
\end{aligned}
$$

7.4   Process of Elimination

Here is another syllogism from Classical Logic in the same vein. Written in binary operator notation, this strong syllogism is expressed as,

$$(\, [\, A \vee B \,] \, \wedge \, \overline{A} \,) \to B$$

If either A or B, or both, are TRUE, and A is FALSE, then B must be TRUE.

In the notation of probability theory, we would write this classical syllogism as $P(B \, | \, \overline{A}, Z)$, the probability that B is TRUE given that we assume that A is FALSE and Z is also TRUE, where Z is the statement that $Z = T \equiv A \vee B$. $P(B \, | \, \overline{A}, Z)$ must then equal 1 to reproduce the result from Classical Logic.

Proceeding as before, we write out Bayes's Theorem for this desired probability,

$$P(B \, | \, \overline{A}, Z) = \frac{P(Z\overline{A}B)}{P(Z\overline{A})}$$

The denominator is,

$$\begin{aligned} P(Z\overline{A}) \;&=\; P(\overline{A} \wedge \, [\, A \vee B \,] \,) \\[2mm] &=\; P(\overline{A}A \vee \overline{A}B) \\[2mm] &=\; P(\overline{A}B) \end{aligned}$$

The numerator is,

$$\begin{aligned} P(Z\overline{A}B) \;&=\; P(B \wedge \, [\, \overline{A}B \,] \,) \\[2mm] &=\; P(B\overline{A}B) \\[2mm] &=\; P(\overline{A}B) \end{aligned}$$

Thus, $P(B \, | \, \overline{A}, Z) = 1$ as expected.

7.4.1   Joint probability table for process of elimination

We'll use a different joint probability table to illustrate the process of elimination. Look at Figure 7.2 at the top of the next page which has different, yet still legitimate, numerical assignments to the joint probabilities indexed by each cell.

The model for assigning the numerical values to the probabilities is different because the logic function being conditioned on has changed. Actually, just a few assignments have been switched from the implication joint probability table shown as Figure 7.1.

The logic function Z has changed in this model from the CONDITIONAL function to the OR function. The new functional assignments of T and F for $f_{15}(A, B)$ are

$$
\begin{array}{cc}
 & Z \\
 & \begin{array}{cc} A & \overline{A} \end{array}
\end{array}
$$

	A	$\overline{A}$	
B	1/4 (Cell 1)	1/4 (Cell 2)	1/2
$\overline{B}$	1/4 (Cell 3)	0 (Cell 4)	1/4
	1/2	1/4	3/4

	A	$\overline{A}$		
B	0 (Cell 5)	0 (Cell 6)	0	1/2
$\overline{B}$	0 (Cell 7)	1/4 (Cell 8)	1/4	1/2
	0	1/4	1/4	
	1/2	1/2		1

(header $\overline{\overline{Z}}$ above second table)

Figure 7.2: *Possible joint probability table reproducing a Classical Logic result, the process of elimination.*

shown. For statement $\overline{Z}$, "It is false that the OR logic function is the model.", I chose $f_2(A, B)$, the NOR logic function. Cell 4, $Z\overline{A}\,\overline{B}$, is 0 because of the presence of the OR function in the conditioning statement. But everything else goes according to plan.

$$
P(B\,|\,\overline{A}, Z) = \frac{P(Z\overline{A}B)}{P(Z\overline{A}B) + P(Z\overline{A}\,\overline{B})}
$$

Because of the 0 in cell 4, $P(Z\overline{A}\,\overline{B}) = 0$ for the second term in the denominator and, as a result,

$$
P(B\,|\,\overline{A}, Z) = \frac{1/4}{1/4 + 0} = 1
$$

7.5 Proof by Cases

Here is another tautology from Classical Logic that is slightly more complicated because it involves three arguments. All of the other examples to this point involved only two arguments. This so–called *proof by cases* [18] is,

$$
(A \vee B) \wedge ((A \rightarrow C) \wedge (B \rightarrow C)) \rightarrow C
$$

This tautology is not too difficult to understand when translated into ordinary English. Supposing that either A or B or both are true, together with the fact that A implies C as well as B implies C, it must be the case that C is true. If both A and B are true, which is possible, then there is no problem. If A weren't true, then B would be true since one of them must be true, and, therefore, B implies that C is true. Likewise, if B weren't true, then A would be true since one of them must be true, and, therefore, A implies that C is true.

We are now using C as a third statement. Nevertheless, Z still stands for the statement that represents the logic function,

$$Z = T \equiv (A \to C) \wedge (B \to C)$$

Therefore, in probability symbol notation we want to find $P(C \mid A \vee B, Z)$ where we hope that this probability equals 1.

We always start with Bayes's Theorem, and then proceed to write out in full the statement Z that is assumed true to the right of the conditioned upon symbol.

$$
\begin{aligned}
P(C \mid A \vee B, Z) &= \frac{P(C \wedge [\, [\, A \vee B \,] \wedge Z \,])}{P([\, A \vee B \,] \wedge Z \,)} \\[2mm]
&= \frac{P(C \wedge [\, [\, A \vee B \,] \wedge [\, [\, A \to C \,] \wedge [\, B \to C \,] \,] \,])}{P([\, A \vee B \,] \wedge [\, [\, A \to C \,] \wedge [\, B \to C \,] \,])} \\[2mm]
&= \frac{P(AC \vee BC)}{P(AC \vee BC)} \\[2mm]
&= 1
\end{aligned}
$$

Obviously, there's quite a bit missing in this proof, namely in the transition from the second to the third step. The following two subsections show an interesting way to fill in the missing steps.

7.5.1   The missing steps in the above proof

Determine the truth table for the function in the denominator. This will tell us what the DNF for the function looks like. A *Mathematica* program determines the assignment of either T or F for all eight possible settings of the variables for the function in the denominator to be,

$$\{\texttt{True, False, True, False, True, False, False, False}\}$$

Thus, the DNF for $[\, A \vee B \,] \wedge [\, [\, A \to C \,] \wedge [\, B \to C \,] \,]$ must be,

$$ABC \vee A\overline{B}C \vee \overline{A}BC$$

This expression can be reduced further. Take out the factor C common to all three terms,

$$C \wedge (AB \vee A\overline{B} \vee \overline{A}B)$$

The three terms in parentheses can be reduced further by taking out the common factor A in the first two terms, resulting in,

$$C \wedge (A \vee \overline{A}B)$$

Now use one of the Absorption properties proved in Chapter Five,

$$(A \vee \overline{A}B) = A \vee B$$

Of course, even easier is the recognition that $AB \vee A\overline{B} \vee \overline{A}B$ is the DNF for the `OR` operator. After this transformation, the **Distributivity** and **Commutativity** axioms yield,

$$C \wedge (A \vee B) = AC \vee BC$$

This gives us what we need for the denominator in Bayes's Theorem. Finally, all that's left to prove is that if we look at the numerator, we obtain

$$C \wedge (AC \vee BC) = ACC \vee BCC = AC \vee BC$$

the same expression as in the denominator. The missing steps in the proof have been filled in.

7.5.2 A curious double–check

There is a curious but satisfying way to double–check this result. Since we are dealing with a Boolean function of three variables, it must be one of Wolfram's three variable cellular automata. Decoding the DNF according to Wolfram's numbering scheme, we see that the rule governing the evolution of the cellular automaton is Rule 168. Wolfram [20] provides a Boolean expression for each one of his 256 elementary one–dimensional cellular automata. The expression given for Rule 168 is, in fact, $(p \vee q) \wedge r$ where Wolfram employs the typical logic nomenclature of p, q, and r for the three variables instead of our $(A \vee B) \wedge C$.

7.5.3 A joint probability table for proof by cases

Just as we have done in the previous examples, we will show a joint probability table to illustrate proof by cases. Look at Figure 7.3 at the top of the next page. This table is larger than the previous examples because it must display four statements instead of three, and thus sixteen cells instead of eight. The top half of the table contains the eight cells for $Z = T$, and the bottom half the eight cells for $Z = F$.

Legitimate numerical values of 0, and numbers less than or equal to 1 (in this case, these numbers are all equal to 1/8) are placed into the 16 cells to capture the uncertainty of the joint statement indexed by each cell. For example, a 0 is entered in cell 3, $P(ZAB\overline{C})$, because C cannot be FALSE if both A and B are TRUE, together with the fact that the logic function representing proof by cases is considered to be the operative model.

The DNF is a heuristic pointing to those cells in the joint probability table which will contain a non–zero assignment. Because of the symmetry involved, the DNF could alternatively be thought of as indicating where the 0 values are to be placed.

Z

	A				Ā			
	B	**B̄**			**B**	**B̄**		
C	1/8 (1)	1/8 (2)	1/4		1/8 (5)	0 (6)	1/8	3/8
C̄	0 (3)	0 (4)	0		0 (7)	0 (8)	0	0
	1/8	1/8	1/4		1/8	0	1/8	3/8

Z̄

	A				Ā			
	B	**B̄**			**B**	**B̄**		
C	0 (9)	0 (10)	0		0 (13)	1/8 (14)	1/8	1/8	1/2
C̄	1/8 (11)	1/8 (12)	1/4		1/8 (15)	1/8 (16)	1/4	1/2	1/2
	1/8	1/8	1/4		1/8	1/4	3/8	5/8

| 1/4 | 1/4 | 1/2 | | 1/4 | 1/4 | 1/2 |

| | | | | 1/2 | 1/2 | | 1.00 |

Figure 7.3: *Possible joint probability table for the proof by cases syllogism.*

In section 7.5.1, the DNF for the denominator was given as,

$$ABC \vee A\overline{B}C \vee \overline{A}BC$$

Hence, cells 1, 2, and 5, under Z, have a numerical assignment of 1/8 while cells 11, 12, 14, 15, and 16, under $\overline{Z}$, have the same numerical assignment. The remaining eight cells contain a 0. The primary constraint relevant to numerical assignments is that the sum is equal to 1. This constraint is indeed satisfied in Figure 7.3.

Some of the marginal sums are listed as well. For example, the marginal sum over the first four cells is $P(ZA) = 1/4$. One of the reasons why a joint probability table is useful resides in the ease of seeing that the sum over cells 1, 2, 3, and 4 is the **Sum Rule** expansion.

$$
\begin{aligned}
P(ZA) &= P(ZABC) + P(ZA\overline{B}C) + P(ZAB\overline{C}) + P(ZA\overline{B}\,\overline{C}) \\
&= 1/8 + 1/8 + 0 + 0 \\
&= 1/4
\end{aligned}
$$

7.5.4 Verifying when C is true

With this background on the joint probability table, subject Bayes's Theorem to a more severe test. Let's take a separate look at the three instances where C must be TRUE as derived from the deductive analysis of proof by cases. In other words, we must confirm the answers given by Classical Logic when we switch over to inferential procedures. To repeat, such an inference relies on the formal manipulation rules of probability theory.

First, assume both A and B TRUE. Place both of these statements, together with the statement concerning the logic function of proof by cases, to the right of the conditioned upon symbol. The statement the information processor is uncertain about is the statement C. It shows up to the left of the conditioned upon symbol. The state of knowledge concerning C given A, B, and Z is $P(C \,|\, A, B, Z)$. Thus,

$$
\begin{aligned}
P(C \,|\, A, B, Z) &= \frac{P(ZABC)}{P(ZAB)} \\[2ex]
&= \frac{P(ZABC)}{P(ZABC) + P(ZAB\overline{C})} \\[2ex]
&= \frac{1/8}{1/8 + 0} \\[2ex]
&= 1
\end{aligned}
$$

Second, we assume just B is TRUE. Here we have a chance to exercise our understanding of marginal sums.

$$
\begin{aligned}
P(C \,|\, B, Z) &= \frac{P(ZBC)}{P(ZB)} \\[2ex]
&= \frac{P(ZBC)}{P(ZBC) + P(ZB\overline{C})} \\[2ex]
&= \frac{P(ZABC) + P(Z\overline{A}BC)}{P(ZABC) + P(Z\overline{A}BC) + P(ZAB\overline{C}) + P(Z\overline{A}B\overline{C})} \\[2ex]
&= \frac{1/8 + 1/8}{1/8 + 1/8 + 0 + 0} \\[2ex]
&= 1
\end{aligned}
$$

$P(ZBC)$ appearing in the numerator is the marginal sum over cells 1 and 5. $P(ZB)$ appearing in the denominator is a marginal sum over cells 1, 5, 3, and 7.

Third, if A is assumed TRUE, then $P(C \,|\, A, Z) = 1$. Since the derivation is the same as just shown for B TRUE, we leave it to the Exercises as practice. Bayes's

Theorem in conjunction with the numerical assignments in the joint probability
table has verified the results arrived at through formal Boolean manipulations on
the logical variables.

7.6   Generalized Syllogisms

So far, all we have done is to verify that probability symbol manipulation does,
in fact, reproduce results from Classical Logic. That is a necessary, but hardly
sufficient, step to support the claim that probability theory *generalizes* Classical
Logic. Now we must show that the same methodology as used above works to
provide sensible answers where Classical Logic fails. Following Jaynes's lead once
again, we begin with what Classical Logic would call an "invalid" argument.

7.6.1   "Invalid" application of *modus ponens*

Consider the *modus ponens* argument again. Instead of asserting A and proving
that B must be true, what if we assert B and wonder about the impact on A? The
binary operator notation for this situation is $B \wedge (A \rightarrow B)$. Classical Logic does
not permit us to say anything at all about A!

Write $P(A \mid B, Z)$ for the probabilistic translation of this conundrum, where Z is
once again the $\rightarrow$ operator. We would be curious to see what *inference*, as opposed
to *deduction*, permits us to say about A.

There is no change in our method of attack. For a change of pace, and as a sort
of double–check, insert the full DNF expression for the implication rather than the
shorter version.

$$P(A \mid B, Z) \;=\; \frac{P(ABZ)}{P(BZ)}$$

$$=\; \frac{P(A \wedge [\, B \wedge [\, AB \vee \overline{A}\, B \vee \overline{A}\,\overline{B} \,]\,])}{P(B \wedge [\, AB \vee \overline{A}\, B \vee \overline{A}\,\overline{B} \,])}$$

Work on the denominator first,

$$P(B \wedge [\, AB \vee \overline{A}\, B \vee \overline{A}\,\overline{B} \,]) \;=\; P(BAB \vee B\overline{A}B \vee B\overline{A}\,\overline{B})$$

$$=\; P(AB \vee \overline{A}\, B)$$

$$=\; P(B \wedge [\, A \vee \overline{A} \,])$$

$$=\; P(B)$$

Now on to the numerator where A is attached to the preceding result via the $\wedge$ operator,

$$P(A \wedge [\, B \wedge [\, AB \vee \overline{A}\, B \vee \overline{A}\, \overline{B} \,] \,] \,) = P(A \wedge B)$$

Substituting these two results back into Bayes's Theorem,

$$P(A \,|\, B, Z) = \frac{P(AB)}{P(B)}$$

When we want to substitute the numerical assignments issuing from some model, it is better to keep the conditioning on Z in the forefront and write,

$$P(A \,|\, B, Z) = \frac{P(AB \,|\, Z)}{P(B \,|\, Z)}$$

as was discussed at the end of the last Chapter.

Substitute the numerical assignments in cells 1 and 2 from the joint probability table shown in Figure 7.1.

$$
\begin{aligned}
P(A \,|\, B, Z) &= \frac{P(AB \,|\, Z)}{P(B \,|\, Z)} \\[2mm]
&= \frac{1/4}{1/4 + 1/4} \\[2mm]
&= 1/2
\end{aligned}
$$

There is no impact on A *under the particular model for implication which assigns numerical values as in Figure 7.1.* The logical implication for A given B and Z is no different than the original marginal sum $P(A) = 1/2$. The information processor is just as uncertain about the statement A when given B as when it was not given this information. At first blush, this outcome seems to be merely another, but essentially equivalent, way to think about the result emanating from a strictly logical analysis.

But understand that this inferential result is conceptually distinct from the deductive one. Deduction in the form of Classical Logic does not even permit the argument to begin! Asserting anything about A given B and implication is an "invalid" logical argument and nothing more can be said. At least the inferential procedure has returned some sort of quantitative assessment. Thus, the information processor knows whether its state of knowledge has changed.

Generalize by inserting different, but still permissible, numerical assignments into the cells of the joint probability table. Doing so will demonstrate that the information processor *can* in fact upgrade its state of knowledge by using the *modus ponens* logic function. Thus, probabilistic inference can circumvent what deduction would label as "invalid."

7.6.2   A joint probability table generalizing *modus ponens*

We return to the relative simplicity of a three statement joint probability table as shown in Figure 7.4. These new numerical values placed into the joint probability table will help us understand the assertions made above about how probability theory generalizes deduction and Classical Logic in a quantitative manner.

	Z A	**Z** $\overline{\text{A}}$			$\overline{\text{Z}}$ A	$\overline{\text{Z}}$ $\overline{\text{A}}$		
B	.02 (Cell 1)	.08 (Cell 2)	.10	**B**	0 (Cell 5)	0 (Cell 6)	0	.10
$\overline{\text{B}}$	0 (Cell 3)	.89 (Cell 4)	.89	$\overline{\text{B}}$	.01 (Cell 7)	0 (Cell 8)	.01	.90
	.02	.97	.99		.01	0	.01	
					.03	.97		1

Figure 7.4: *Possible joint probability table for a generalization of* modus ponens.

In the last section we determined that,

$$P(A \mid B, Z) = \frac{P(AB \mid Z)}{P(B \mid Z)}$$

Substituting the numerical values from Figure 7.4

$$P(A \mid B, Z) = \frac{0.02}{0.02 + 0.08}$$

$$= 0.20$$

Thus, conditioning on the knowledge that B is true and similarity to the implication $A \to B$, has raised the degree of belief in A from 0.03 to 0.20. A is not certain, but it is now more certain after B has happened.

Contrary to the first example, $P(A \mid B, Z)$ is now greater than $P(A)$. The only conclusion is that the information processor's state of knowledge has changed. You may recall that Classical Logic warned us not to even attempt to argue about A from an implication model. On the other hand, inference showed us how much we could change our state of knowledge about A via probability as some measure of the "logical implication" of B's truth.

7.6.3 Life on Mars

Here is a little example to provide some more support for what has so far just been a purely syntactical development. Let's say that L is the statement, "Life, as we know it, exists on Mars." W is the statement, "Liquid water is available on Mars."

Furthermore, we are tentatively exploring the ramifications of the implication model, $L \rightarrow W$. What are the logical implications on an information processor's belief in life on Mars if liquid water has been discovered? In symbols, what is $P(L \,|\, W, L \rightarrow W)$?

Using deduction and Classical Logic we can say nothing after this discovery. Had life been found on Mars, we would have been justified in asserting the presence of water under the current model, but, as luck would have it, not the other way around. Naturally, we have phrased the problem in exactly the same way as presented above when we were curious whether an inference about A could be made based on the fact that B had occurred.

Using the same numerical assignments given in Figure 7.4, the initial assignment said that one legitimate probability for life on Mars under the current model was $P(L) = P(A) = 0.03$. The current scientific evidence led to an initial assignment of $P(W) = P(B) = 0.10$ for the presence of liquid water available to life on Mars.

Assuming that liquid water has actually been detected on Mars, the formal manipulation rules in the guise of Bayes's Theorem permit us to increase the probability of life on Mars from its previous value of 0.03 to a current value of $P(L \,|\, W, L \rightarrow W) = 0.20$. Our state of knowledge has thus changed to being more certain about life due to a known observation and the tentative acceptance of some model.

If it should ever turn out that life as we know it was indeed present on Mars, but there was no liquid water available to that life form, then the current implicational model is FALSE. This is the bail–out represented in cell 7 by $P(\overline{Z}A\overline{B}) = 0.01$.

7.6.4 Jaynes's version

Jaynes presented a version different from our development. He relied on the more typically seen conditional probabilities rather than our preferred joint probabilities. Applying the product rule to the numerator of Bayes's Theorem and conditioning throughout on the model Z, we have,

$$P(A \,|\, B, Z) = \frac{P(B \,|\, A, Z)\, P(A \,|\, Z)}{P(B \,|\, Z)}$$

But if A is assumed true in the conditional probability appearing as the first term of the numerator on the right hand side, then $P(B \,|\, A, Z)$ must be 1 from the implication $Z = T \equiv A \rightarrow B$. This results in,

$$P(A \,|\, B, Z) = \frac{P(A \,|\, Z)}{P(B \,|\, Z)}$$

Checking this derivation with the numerical assignment for the Mars example we find that,

$$P(A \mid B, Z) = \frac{0.02}{0.10} = 0.20$$

confirming our previously derived result. Since Z was a given for both probabilities $P(A \mid Z)$ and $P(B \mid Z)$, the marginal sums for $P(A) = 0.02$ and $P(B) = 0.10$ were selected from the separate 2×2 table on the left hand side of the full joint probability table.

So far, we have always placed $Z = T$ to the right of the conditioned-upon symbol to indicate that a particular logic function was operative. Or, in other words, that a particular model was making the numerical assignments from a joint probability table. What if we wrote $P(B \mid A)$ where Z is missing? How do we take account of the models that are not specified? This point is taken up further in an exercise.

Returning to Jaynes, he directs our attention to the constraints imposed on probabilities. Since $P(A \mid Z) \leq P(B \mid Z)$ and $P(B \mid Z) \leq 1$, it must be that $P(A \mid B, Z) \geq P(A \mid Z)$. Thus, we have learned that, although A is not raised to certainty by any means given that B and Z are true, its probability has either stayed the same, or has been raised from $P(A \mid Z)$ where B was not known. We have already experienced the equality part of this relationship when we found out that,

$$P(A \mid B, Z) = P(A) = 1/2$$

for the numerical assignment imparted by the model in Figure 7.1.

Rather than being an "invalid" argument, probability generalizes Classical Logic by showing us quantitatively how much we should increase our degree of belief that A is true if, in fact, B happens and the implication $A \to B$ is true. We'll discuss a similar "invalid" use of this Classical Logic function next.

7.6.5   Another "invalid" application of *modus ponens*

What about another similar situation that cannot be handled by Classical Logic? We have the same statement Z as before, the implication $A \to B$, but now A is FALSE. What is the impact of this information on B?

Writing $P(B \mid \overline{A}, Z)$ for the probabilistic translation of this mystery, we see what Bayes's Theorem has to say. In essence, we want to observe the difference between the inferential and deductive approaches to the processing of information. We revert back to inserting the shorter expression for the logical operator of implication.

$$P(B \mid \overline{A}, Z) = \frac{P(B\overline{A}Z)}{P(\overline{A}Z)}$$

$$= \frac{P(B \wedge [\,\overline{A} \wedge [\,\overline{A} \vee B\,]\,])}{P(\overline{A} \wedge [\,\overline{A} \vee B\,])}$$

$$P(\overline{A} \wedge [\,\overline{A} \vee B\,]\,) = P(\overline{A}\,\overline{A} \vee \overline{A}\,B)$$

$$= P(\overline{A} \vee \overline{A}\,B)$$

$$= P(\overline{A})$$

$$P(B \wedge \overline{A} \wedge [\,\overline{A} \vee B\,]\,) = P(B \wedge \overline{A})$$

$$P(B\,|\,\overline{A}, Z) = \frac{P(\overline{A}B)}{P(\overline{A})}$$

Let's look at Jaynes's argument once again to get a sense of the direction in which the probabilities must change. Applying the product rule to the numerator, we have,

$$P(B|\overline{A}, Z) = \frac{P(\overline{A}\,|\,B, Z)\,P(B\,|\,Z)}{P(\overline{A}\,|\,Z)}$$

But we just learned in the previous section that $P(A\,|\,B, Z) \geq P(A\,|\,Z)$. Therefore, since $P(\overline{A}\,|\,B, Z) = 1 - P(A\,|\,B, Z)$, $P(\overline{A}\,|\,B, Z) \leq P(\overline{A}\,|\,Z)$. And thus, the fraction multiplying $P(B\,|\,Z)$ must be less than or equal to 1. Our final goal is then reached by writing,

$$P(B|\overline{A}, Z) \leq P(B\,|\,Z)$$

Our belief in B stays the same or is lessened once we have learned that A did not happen and given that the implication $Z \equiv A \to B$ is assumed true.

We can verify this last result with a numerical example using the joint probability table given for the Life on Mars scenario. Moreover, we will discover the exact quantitative consequences on our state of knowledge about the existence of water after finding out that life does not exist on Mars.

Writing out Bayes's Theorem in the form conducive to joint probabilities,

$$P(B\,|\,\overline{A}, Z) = \frac{P(Z\overline{A}B)}{P(Z\overline{A}B) + P(Z\overline{A}\,\overline{B})}$$

Substituting the numerical values from the appropriate cells in the joint probability table, we have,

$$P(B\,|\,\overline{A}, Z) = \frac{0.08}{0.08 + 0.89} = 0.0825$$

Thus, our degree of belief in B is indeed lessened to the tune of 0.0825 when A does not occur. Compare this to $P(B\,|\,Z) = 0.10$ when A was not known. Translated to the Mars scenario, if life, as we know it, definitely does not exist on Mars, then we become slightly more certain that liquid water is not available either.

7.7   A Criminal Inference

These past few examples have not only extended Classical Logic. They have opened our eyes to the idea of *inference*. At the highest level, what we are interested in is how our belief (state of knowledge) in some statement changes (is updated) when we condition on some actual events that *have* happened, together with some model of how the world works. In some of the previous examples, logical implication was employed as a stand–in for these kinds of tentative ontological models.

To provide a concrete mental image for the upcoming symbol manipulations, consider a make–believe criminal scenario. Suppose that some credible model of murder asserts that if suspect A commits a murder, then there is blood on the suspect's clothes (this is statement B). In addition, if suspect A has committed a murder, then the suspect will take the victim's car (this is statement C). Taken together, the deductive model Z states that A implies B and A also implies C.

This tentative model of how the world of criminality works is summarized as the statement,

$$Z \equiv (A \to B) \land (A \to C)$$

Now let's suppose that a murder has been committed and suspect A has the victim's blood on his or her clothes. Thus, B is true. Unfortunately for A, he or she also happens to have been seen driving the victim's car, so C is true as well. We want to update our state of knowledge about whether the suspect committed the murder. For a quantitative solution, we state this probabilistically as $P(A \mid B, C, Z)$.

In other words, we admit that we cannot realistically claim with any certainty, that is, deduce by logical reasoning, whether the suspect murdered the victim. However, we *can* do the next best thing and use the information that we do have, the facts and a close analog of the above logic function, to create a model of criminal behavior, in order to change our degree of belief about the suspect.

In fact, this example is just an extension of the "invalid" *modus ponens* scenario. We don't know that A is true, and therefore can't use logic in the forward direction to deduce B and C. What we do know is that B and C are true, and we wonder what the consequences might be for believing in A.

Everything proceeds as before. We phrase our inquiry in the form of Bayes's Theorem. Then, in Exercise 7.9.11, we work out the details for the numerator and denominator in symbolic terms using the allowable transformations from Boolean Algebra. Initially, we arrive at the answer through this rather long and convoluted route. But the same answer can be found using numerical assignments to a joint probability table. This latter approach seems more amenable for a general attack on inferential problems.

In the end, not too surprisingly, the symbolic answer looks like this:

$$P(A \mid B, C, \mathcal{M}_k) = \frac{P(ABC \mid \mathcal{M}_k)}{P(ABC \mid \mathcal{M}_k) + P(\overline{A}BC \mid \mathcal{M}_k)}$$

We have switched to a notation where a model $\mathcal{M}_k$ assigning legitimate numerical values to the joint probabilities is given instead of some logic function Z. In future work, we want to exploit our newly found freedom to deviate from the strict shackles of a logic function as indicated by Z. It's another way to highlight the consequences of an inferential, as contrasted to a purely deductive, approach.

7.7.1   A joint probability table for numerical examples

Despite our emphasis on strictly symbolic results, it is helpful to flesh out results of this kind with some numbers just to get a sense of what is involved. Working through such a numerical exercise also serves to remind us that our ultimate goal is not symbology manipulation as some soulless kind of formal game, but rather the study of manipulation rules to assist us in making inferences.

Do not take the numbers in this exercise too seriously. They are there mainly to show how a state of knowledge about some statement can be updated in a quantitative manner via Bayes's Theorem.

In addition, this numerical example serves the very useful purpose of allowing us to double–check any symbolic results. There is always a lingering suspicion after some long train of operations that a mistake might have gone undetected along the way. If we set up a joint probability table, it is easy to see how various joint statements are assembled from more elementary joint statements.

To that end, inspect Figure 7.5 below for a possible joint probability table of the criminal scenario with some numbers assigned as legitimate probabilities. The actual numbers stem from the *information* in a given model.[4] The motivation for such a model is the tentative working hypothesis embodied in the two implications about murderers, blood, and cars.

Figure 7.5: *Possible joint probability table for solving a numerical exercise about criminal behavior.*

[4]This assignment of numbers falls under the purview of the second half of probability theory, that is, from the *Maximum Entropy Principle* and *Information Geometry*.

The joint probability table is constructed from only the three statements A, B, and C and consists of eight cells. Heretofore, we have emphasized the logic function's role as a model by explicitly including Z and $\overline{Z}$ as part of the joint probability table. Should we not have shown a 16 cell table to include both possibilities under Z and some $\overline{Z}$?

This joint probability table is condensed because we are considering just the one situation where $Z = T$. In effect, what we are doing here is considering just *one* model $\mathcal{M}_k$ that assigns legitimate numerical values to all of the joint probabilities. Previously, we had introduced two models, one where $Z = T$ and one where $Z = F$. This topic concerning model specification is taken up in much greater detail in the succeeding Chapters.

Despite this simplification, we have managed to generalize our inferences with this joint probability table. If we had adhered strictly to the implicational model of $Z \equiv (A \rightarrow B) \wedge (A \rightarrow C)$, then there would have been 0s placed in the appropriate cells of the table as we have done in all the previous examples. For example, during the symbol manipulation proof, we discovered that cell 3 in the table, that is, the cell indexing the uncertainty about the joint statement, "A and B are both TRUE, but C is FALSE," was a cell where a 0 should be placed.

Now, however, positive numbers are inserted as legitimate numerical assignments to every joint probability indexed by the table. We are generalizing away from the strictly implicational model. As can be observed, a very small numerical assignment is now inserted in cell 3 instead of a 0. It would not be unfair to say that we are investigating an "almost pure implicational model." To highlight this departure from our previous way of doing business, Z will be replaced by the notation for a model, $\mathcal{M}_k$.

Reflect on the overall plausibility of this table. The highest probability is given to the joint statement, $P(\overline{A}\,\overline{B}\,\overline{C}) = 0.94$, for any given suspect not being the murderer, along with not having any of the victim's blood on their clothing, nor in possession of the victim's car. All of the other joint statements are relatively low in the probability that remains after this most probable case has been covered. $P(\overline{A}\,\overline{B}C) = 0.02$ covers the case where a suspect is not the murderer and isn't covered in the victim's blood, but nevertheless was driving the victim's car. The victim might have lent the car to an innocent neighbor.

And, it is relatively unlikely, although not strictly ruled out, that the suspect did not murder the victim despite having the incriminating evidence of both the victim's blood and the victim's car, $P(\overline{A}BC) = 0.005$. Perhaps, the innocent suspect had found the victim by the side of the road, carried the unfortunate person, dripping in blood, to the victim's car, and then drove to the nearest hospital.

When the suspect is the murderer, the largest probability, $P(ABC) = 0.01$, although still low in relative terms, is allocated to the suspect having both pieces of incriminating evidence. The lowest probability of all, $P(A\,\overline{B}\,\overline{C}) = 0.002$, is given

(always through the auspices of some model) to the case where the suspect did, in fact, murder the victim, but was clever enough to rid himself of all incriminating evidence.

The joint probability table also provides all of the marginal probabilities. The marginal probability for anyone, murderer or not, driving around in the victim's car is $P(C) = 0.038$.

But now let's move on to question posed at the outset. What is the revised probability (the updated state of knowledge) that the suspect is the murderer when it has become a fact that the suspect is in possession of the victim's car, and the victim's blood is all over the suspect's clothing?

The information processor is then faced with the overriding inferential question: What is the revised probability that some suspect is the murderer given that both pieces of incriminating evidence have been found true? The IP was faced with an initial marginal probability from this model saying that $P(A \mid \mathcal{M}_k) = 0.02$. How is this number changed after we find the answer to $P(A \mid B, C, \mathcal{M}_k)$?

Notice that the model has now been made clearly visible as something known by placing it to the right of the conditioned upon symbol. This is imperative when we are not discussing abstract probability symbols, but have moved on to the actual numerical values dictated by some model.

After substituting the numerical assignments from the appropriate cells of the joint probability table, we find that,

$$
\begin{aligned}
P(A \mid B, C, \mathcal{M}_k) &= \frac{P(ABC \mid \mathcal{M}_k)}{P(BC \mid \mathcal{M}_k)} \\[2mm]
&= \frac{P(ABC \mid \mathcal{M}_k)}{P(ABC \mid \mathcal{M}_k) + P(\overline{A}BC \mid \mathcal{M}_k)} \\[2mm]
&= \frac{0.010}{0.010 + 0.005} \\[2mm]
&= 0.67
\end{aligned}
$$

We revise our state of knowledge that the suspect is the murderer given the presence of the incriminating evidence upwards from $P(A \mid \mathcal{M}_k) = 2\%$ to,

$$P(A \mid B, C, \mathcal{M}_k) = 67\%$$

As always, we have to issue the caveat that such a quantitative inference depends not only on known facts like B and C, but also on the model $\mathcal{M}_k$ that is "almost like" $Z \equiv (A \rightarrow B) \wedge (A \rightarrow C)$.

7.8   Circumstantial Evidence

Continue this line of thinking. What if there were a whole series of consequences if A were true? Let $\mathcal{M}_k$ stand for the tentative working hypothesis, or model, that all the following implications hold,

$$A \quad \to \quad B$$

$$A \quad \to \quad C$$

$$A \quad \to \quad D$$

$$\vdots$$

$$A \quad \to \quad \cdots$$

What is the revised probability for A if all of these events have occurred? Note that we cannot claim anything for certain because this is not a problem that falls within the domain of Classical Logic. Rather, we face a problem of inference. It might be called the problem of "overwhelming circumstantial evidence."

The answer is quite easy. Just employ the rule of probability manipulation encapsulated in Bayes's Theorem,

$$P(A \,|\, B, C, D, \cdots, \mathcal{M}_k) = \frac{P(ABCD\cdots \,|\, \mathcal{M}_k)}{P(ABCD\cdots \,|\, \mathcal{M}_k) + P(\overline{A}BCD\cdots \,|\, \mathcal{M}_k)}$$

Now the actual numerical probabilities for complicated statements like these are going to be very small. After all, the total probability of 1 has to be divided up among a lot of different situations. However, the crucial numbers revolve around the relative probability of A's occurrence or non–occurrence given the "incriminating evidence," B, C, D, and so on, that, however improbably, has taken place.

If an assigned probability for the joint occurrence of $ABCD\cdots$ is higher than the probability assigned to the joint occurrence of $\overline{A}BCD\cdots$, then, despite the fact that these joint probabilities may be very small, the revised probability about A may become quite substantially large. For example, the joint occurrence of $ABCD\cdots$ might be the very small legitimate probability of 0.0001 and the joint occurrence of $\overline{A}BCD\cdots$ might be somewhat smaller at 0.00005, but then our degree of belief that A actually did occur is raised quite dramatically to $P(A \,|\, B, C, D, \cdots, \mathcal{M}_k) = 2/3$.

Our criminal justice system does not seem to take overwhelming circumstantial evidence very seriously. The judicial mind–set is locked into reasoning forward by deduction and Classical Logic. But, clearly, probability generalizes deduction and permits justifiable inferences to be made on the basis of the available evidence. Laplace argued for the use of probability and Bayes's Theorem in criminal matters over 200 years ago and we know what kind of reception that received. Nothing has changed in the interim. We would all be better served under the realization that inferences are an inevitable part of social justice.

7.9   Solved Exercises for Chapter Seven

Exercise 7.9.1. The implication logic function was treated extensively in this Chapter. Carry out the same kind of analysis for $P(B \mid A, Z)$ where the given model Z is the logic function labeled as $f_{14}(A, B)$.

Solution to Exercise 7.9.1

The implication logic function was labeled as $f_{13}(A, B)$ in Chapter Two. It was given the operator symbol $\rightarrow$ and the name CONDITIONAL. Thus, we could write $f_{13}(A, B) \equiv A \rightarrow B$. Another one of the sixteen logic functions was $f_{14}(A, B)$. Similarly, we wrote this function as $f_{14}(A, B) \equiv A \leftarrow B$. There was a reason for giving it the operator symbol $\leftarrow$. The name CONDITIONAL was given to f_{14} because it represented the reverse implication from B to A.

Let the model be Z, that is, where the logic function $A \leftarrow B$ dictates the numerical assignments. The second model $\overline{Z}$ will be $A \star B$, or $f_3(A, B)$.

Recall that the DNF expansion of $f_{14}(A, B)$ is $AB \vee A\overline{B} \vee \overline{A}\,\overline{B}$ because T is assigned as the functional value at the following three settings of the A and B variables: (1) $A = T, B = T$, (2) $A = T, B = F$, and (3) $A = F, B = F$. Compare this with the DNF expansion of $f_{13}(A, B)$ as $AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$. The only difference is in the second term where the negation is swapped from B to A. The upshot is that $f_{14}(A, B)$ acts as an implication function in the opposite direction, $B \rightarrow A$.

Bayes's Theorem relies on substituting the DNF expansion for $Z = T$ together with the formal manipulation rules from Boolean Algebra,

$$P(B \mid A, Z) = \frac{P(BAZ)}{P(AZ)}$$

$$P(BAZ) = P(B \wedge [\, A \wedge [\, AB \vee A\overline{B} \vee \overline{A}\,\overline{B}\,]\,]\,)$$

$$= P(B \wedge [\, AAB \vee AA\overline{B} \vee A\overline{A}B\,]\,)$$

$$= P(B \wedge [\, AB \vee A\overline{B}\,]\,)$$

$$= P(BAB \vee BA\overline{B})$$

$$= P(AB)$$

$$P(AZ) = P(A \wedge [\, AB \vee A\overline{B} \vee \overline{A}\,\overline{B}\,]\,)$$

$$= P(AB) + P(A\overline{B})$$

$$P(B \mid A, Z) = \frac{P(AB)}{P(AB) + P(A\overline{B})}$$

$$P(B \,|\, A, Z) = \frac{P(AB \,|\, Z)}{P(A \,|\, Z)}$$

$f_{14}(A, B)$, just like the regular implication function, as well as the other two logic functions which take on the functional value of T at three variable settings, can be described with a joint probability table with 0 located in one cell (when $Z = T$) and legitimate values greater than 0 in the other three cells (when $Z = T$).

Establish a mirror symmetry with 0s in three cells and a legitimate value in the remaining cell when $Z = F$. The DNF expansion for the function in question tells us where these three cells are located. The numerical assignment of $1/4$ appears in the relevant cells in the joint probability table for the three variables as shown in Figure 7.6.

Figure 7.6: $2 \times 2 \times 2$ *joint probability table to illustrate logic function* $f_{14}(A, B)$.

Substituting these numerical assignments into Bayes's Theorem, we find that,

$$P(B \,|\, A, Z) = \frac{P(AB \,|\, Z)}{P(A \,|\, Z)} = \frac{1/4}{1/4 + 1/4} = 1/2$$

If A is TRUE and $A \leftarrow B$ is the model, then B is logically implied as TRUE with measure of $1/2$. Contrast this with the certain implication of A TRUE when B is TRUE and $A \leftarrow B$ is the model, $P(A \,|\, B, A \leftarrow B) = 1$.

Exercise 7.9.2. What are the other two logic functions besides "→" and "←" mentioned above that have a functional assignment of T for three variable settings?

Solution to Exercise 7.9.2

The other two functions, from the total of four functions which take on the functional value of T at three of the four variable settings, are the NAND operator, $\uparrow$, and the OR

operator, $\vee$. Joint probability tables for these two functions will exhibit the same general pattern as already seen for the two CONDITIONAL operators.

The DNF for NAND is $A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B}$, so a 0 will appear for $P(ZAB)$ in cell 1. Legitimate numerical values greater than 0 for $P(Z\overline{A}B)$, $P(ZA\overline{B})$, and $P(Z\overline{A}\,\overline{B})$ will appear in cells 2, 3, and 4. Likewise, 0s will appear in cells 6, 7, and 8, $P(\overline{Z}\,\overline{A}B)$, $P(\overline{Z}A\overline{B})$, and $P(\overline{Z}\,\overline{A}\,\overline{B})$. A legitimate numerical value will appear in cell 5, $P(\overline{Z}AB)$.

The DNF for OR is $AB \vee A\overline{B} \vee \overline{A}B$, so a 0 will appear for $P(Z\overline{A}\,\overline{B})$ in cell 4. Legitimate numerical values greater than 0 for $P(ZAB)$, $P(Z\overline{A}B)$, and $P(ZA\overline{B})$ will appear in cells 1, 2, and 3. Likewise, 0s will appear in cells 5, 6, and 7, $P(\overline{Z}AB)$, $P(\overline{Z}\,\overline{A}B)$, and $P(\overline{Z}A\overline{B})$. To conclude, we can say that a legitimate numerical value will appear in cell 8, $P(\overline{Z}\,\overline{A}\,\overline{B})$.

Exercise 7.9.3. Describe a discernible pattern among the sixteen logic functions based on the first two exercises.

Solution to Exercise 7.9.3

The four functions taking on the functional assignment of T at three of the four possible variable settings, that is, models acting as Z,

1. f_{12}, (NAND, $\uparrow$),

2. f_{13}, (CONDITIONAL, $\rightarrow$),

3. f_{14}, (CONDITIONAL, $\leftarrow$), and

4. f_{15}, (OR, $\vee$)

are balanced off by the four functions taking the functional assignment of F at the same variable settings, that is, models acting as $\overline{Z}$,

1. f_5, (AND, $\wedge$),

2. f_4, (DIFFERENCE, $\diamond$),

3. f_3, (DIFFERENCE, $\star$), and

4. f_2, (NOR, $\downarrow$)

Furthermore, the three functions where the functional assignment of T is made at two of the variable settings are balanced off by three other functions where the functional assignment of F is made at exactly the same place. So far we have mentioned fourteen of the sixteen logic functions. The final two, f_{16} and f_1, are balanced off by having T as the functional assignment for all variable settings where the other has F as the functional assignment at all four variable settings.

Exercise 7.9.4. Show how probability theory reproduces the classical result for *reductio ad absurdum*.

Solution to Exercise 7.9.4

The *reductio ad absurdum* argument from Classical Logic is perhaps one of the strangest and most controversial. It is the infamous "law of the excluded middle" that so bedeviled the Dutch mathematician L.E.J. Brouwer. Nevertheless, we can show how probability theory reproduces this classical tautology using the same techniques described in this Chapter.

Reductio ad absurdum is a logic function with two arguments,

$$Z \equiv (\overline{A} \to B) \wedge (\overline{A} \to \overline{B})$$

In words, it says that if the negation of a statement A implies that another statement B is TRUE and, furthermore, that the negation of that same statement A implies that B is FALSE, then the original non–negated statement A must be true. If some statement $\overline{A}$ implies a contradiction, then its falsification must be true.

Couched in the notation of probability, we wonder if $P(A \mid Z) = 1$? Apply Bayes's Theorem to this conditional probability,

$$P(A \mid Z) = \frac{P(ZA)}{P(Z)}$$

Carry out the expansion and subsequent reduction for the denominator according to the axioms and derived theorems of Boolean Algebra,

$$
\begin{aligned}
P(Z) &= P(\,[\,\overline{A} \to B\,] \wedge [\,\overline{A} \to \overline{B}\,]\,) \\[2mm]
&= P(\,[\,A \vee B\,] \wedge [\,A \vee \overline{B}\,]\,) \\[2mm]
&= P(AA \vee A\overline{B} \vee BA \vee B\overline{B}) \\[2mm]
&= P(A \vee A\overline{B} \vee AB) \\[2mm]
&= P(A \vee (A \wedge (\overline{B} \vee B))) \\[2mm]
&= P(A \vee A) \\[2mm]
&= P(A)
\end{aligned}
$$

Next, work out the now easy numerator,

$$P(ZA) = P(A \wedge A) = P(AA) = P(A)$$

Thus, the rules of probability manipulation confirm that A is certain,

$$P(A \mid Z) = \frac{P(A)}{P(A)} = 1$$

Exercise 7.9.5. Provide a joint probability table to numerically verify the above result.

Solution to Exercise 7.9.5

Figure 7.7 presents one possibility.

Figure 7.7: $2 \times 2 \times 2$ *joint probability table to illustrate* reductio ad absurdum.

Substitute the numerical assignments from this joint probability table to confirm that A is certain given *reductio ad absurdum*.

$$
\begin{aligned}
P(A \,|\, Z) &= \frac{P(ZA)}{P(Z)} \\[2mm]
&= \frac{P(ZAB) + P(ZA\overline{B})}{P(Z)} \\[2mm]
&= \frac{1/4 + 1/2}{3/4} \\[2mm]
&= 1
\end{aligned}
$$

Furthermore, we see that A is certain no matter what B is. Whether A is conditioned on B's truth or falsity, or whether A is independent of B, the outcome is all the same, A is certain to be true.

$$
P(A \,|\, Z) = P(A \,|\, B, Z) = P(A \,|\, \overline{B}, Z) = 1
$$

For example,

$$P(A \,|\, \overline{B}, Z) \;=\; \frac{P(ZA\overline{B})}{P(Z\overline{B})}$$

$$=\; \frac{P(ZA\overline{B})}{P(ZA\overline{B}) + P(Z\overline{A}\,\overline{B})}$$

$$=\; \frac{1/2}{1/2 + 0}$$

$$=\; 1$$

It is important to keep in mind the distinction between joint, marginal, and conditional probabilities. $P(ZA) = P(ZAB) + P(ZA\overline{B})$ is a marginal probability composed from two joint probabilities. It is the sum of the assignments in cell 1 and cell 3. Thus, $P(ZA) = 3/4$. Also,

$$P(Z) = P(ZAB) + P(Z\overline{A}B) + P(ZA\overline{B}) + P(Z\overline{A}\,\overline{B}) = 3/4$$

It is the marginal sum of the assignments to the four joint probabilities in cells 1 through 4. But using **Commutativity** followed by the **Product Rule**, the marginal probability can also be decomposed into a term involving the conditional probability,

$$P(ZA) = P(AZ) = P(A \,|\, Z)\, P(Z) = 1 \times 3/4 = 3/4$$

Figure 7.7 showed a joint probability table consisting of eight cells. Eight joint statements were constructed from the three variables A, B, and Z with each variable taking on only two values T or F. Under $Z = T$, the left sub–table, we have the logic function $f_{11}(A, B)$ whose DNF expansion dictates that a 0 must be placed into cell 2, $Z\overline{A}B$, and cell 4, $Z\overline{A}\,\overline{B}$. These two cells are where the DNF expansion of $A \triangleleft B$ has a coefficient of F. Conversely, the DNF expansion of $A \triangleleft B$ has a coefficient of T for ZAB and $ZA\overline{B}$, and so some legitimate numerical assignment must be inserted into cells 1 and 3.

$Z = T$ stands for some deductive model representing $A \triangleleft B$, which we gave the label as the A operator, while $Z = F$, the right sub–table, stands for a second, "dual" deductive model. This second model is the mirror image of $A \triangleleft B$. Its DNF expansion has a T where $A \triangleleft B$ had an F, and *vice versa*. This second model is, in fact, $f_6(A, B)$, or $A \vdash B$, called the NOT A operator.

Here is an interesting quote from Roger Penrose about this classical syllogism:[5]

> "... one of the most time–honoured and fruitful principles ever to be put forward in mathematics—very possibly first introduced by the Pythagoreans—
> [is] called *proof by contradiction* (or *reductio ad absurdum* to give it its Latin

---

[5] *The Road To Reality*, Roger Penrose, pp. 42–43

name). According to this procedure, in order to prove that some assertion is
true, one first makes the supposition that the assertion in question is *false*,
and then one argues from this that some contradiction ensues. Having found
such a contradiction, one deduces that the assertion must be true after all."

It is my contention that showing the joint probability tables for logic functions
reduces their mystery. This technique, I argue, makes concepts initially found to
be difficult more understandable because these concepts are interpreted within the
wider context of probability as a generalization of logic.

Exercise 7.9.6. Employ *Mathematica* (after reading the Appendices) to verify *reductio ad absurdum* in yet another way.

Solution to Exercise 7.9.6

Since *reductio ad absurdum*,

$$Z \equiv (\overline{A} \to B) \wedge (\overline{A} \to \overline{B})$$

is simply a logic function of two variables, we can use *Mathematica* to write Z as,

And[Implies[Not[p], q], Implies[Not[p], Not[q]]]

Then, $f_{11}(A, B)$, or $A \triangleleft B$ is written as **f11 = BooleanFunction[12, 2]**. Set
up the proposed tautology between *reductio ad absurdum* and $f_{11}(A, B)$ as,

logicExpression[p_, q_] := Equivalent[And[Implies[Not[p], q],
Implies[Not[p], Not[q]]], f11[p, q]]

where **Equivalent[arg1, arg2]** implements the logical equivalence between the
two arguments. Evaluating **TautologyQ[logicExpression[p, q]]** returns
True verifying that *reductio ad absurdum* has the same functional assignments as
one of the 16 logic functions, namely $f_{11}(A, B)$, or $A \triangleleft B$ and called the binary
operator **A**. *Reductio ad absurdum* is then true to its name because it does reduce
down to the triviality that $P(A \mid A) = 1$.

Exercise 7.9.7. Is there some connection between the *modus ponens* example and the ordering relationship in Boolean Algebra?

Solution to Exercise 7.9.7

In fact, the generic ordering relationship $a \leq b$ from Boolean Algebra discussed in
Chapter Four, is the same as logical implication $A \to B$.

One of the distinguishing characteristics of the ordering relationship was,

$$a \circ b' = F$$

which is reflected in the numerical assignment of $P(A\overline{B}) = 0$. From the joint probability table implementing the implication shown as Figure 7.1, we certainly have,

$$P(A) \leq P(B) \equiv P(AB) + P(A\overline{B}) \leq P(AB) + P(\overline{A}B)$$

Since $P(A\overline{B}) = 0$,

$$P(AB) \leq P(AB) + P(\overline{A}B)$$

$P(AB)$ must certainly be less than or equal to itself because $P(\overline{A}B)$ as a probability cannot be negative.

Another distinguishing characteristic of the ordering relationship in Boolean Algebra was $a' \bullet b = T$. This arises when we complement both sides of the $a \circ b' = F$ equation and then use **De Morgan's axiom**,

$$(a \circ b')' = (F)' \equiv a' \bullet b = T$$

The analogous probability expression is $P(\overline{A} \vee B) = 1$. But,

$$P(\overline{A} \vee B) = P(\overline{A}) + P(B) - P(\overline{A}B)$$

The marginal sum for $P(\overline{A})$ is $P(\overline{A}B) + P(\overline{A}\,\overline{B})$. The marginal sum for $P(B)$ is $P(AB) + P(\overline{A}B)$.

$$P(\overline{A} \vee B) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(AB) + P(\overline{A}B) - P(\overline{A}B)$$

$$= P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(AB)$$

But if,

$$P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(AB) = 1$$

then this is the same as the model for $A \rightarrow B$ since $P(A\overline{B}) = 0$.

Exercise 7.9.8. Prove that which was so cavalierly asserted in the last paragraph of the previous exercise.

Solution to Exercise 7.9.8

From the rule for distributing the probability operator, we have,

$$P(\overline{A} \vee B) = P(\overline{A}) + P(B) - P(\overline{A}B)$$

Now expand $\overline{A}$ and B by the ruse of "multiplying by 1,"

$$T \wedge \overline{A} = \overline{A}$$

$$B \vee \overline{B} = T$$

$$(B \vee \overline{B}) \wedge \overline{A} = \overline{A} \wedge (B \vee \overline{B})$$

$$\overline{A} = \overline{A}B \vee \overline{A}\,\overline{B}$$

B is expanded in exactly the same fashion.

$$T \wedge B = B$$

$$T = A \vee \overline{A}$$

$$B = (A \vee \overline{A}) \wedge B$$

$$= AB \vee \overline{A}B$$

Substituting these expansions for $\overline{A}$ and B, we have,

$$P(\overline{A} \vee B) = P(\overline{A}B \vee \overline{A}\,\overline{B}) + P(AB \vee \overline{A}B) - P(\overline{A}B)$$

Once again use the rule for distributing the probability operator over $\vee$ in the first two terms. The cross–product is always F so that,

$$P(\overline{A} \vee B) = P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(AB) + P(\overline{A}B) - P(\overline{A}B)$$

$$= P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(AB)$$

All of this is contingent upon the acceptance of the model $Z = T$. And, of course, this model is the implication logic function. Generically, we know that Bayes's Theorem allows us to write,

$$P(E \mid Z) = \frac{P(ZE)}{P(Z)}$$

where E is some Boolean expression. So, if $P(ZE) \equiv P(Z\overline{A}B) + P(Z\overline{A}\,\overline{B}) + P(ZAB)$, then ultimately,

$$P(\overline{A} \vee B) = \frac{1/4 + 1/4 + 1/4}{3/4} = 1$$

where the numerical assignment from Figure 7.1 is used for convenience and $P(Z)$ can be read off as the marginal sum over the first four cells of the joint probability table.

Exercise 7.9.9. In section 7.5.2, it was mentioned that Wolfram's Rule 168 for cellular automata provided a curious double–check on the results. Decode Rule 168 to find the DNF and then express it in *Mathematica* syntax.

Solution to Exercise 7.9.9

Decoding Rule 168 we find that,

$$168 = 128 + 32 + 8 = 2^7 + 2^5 + 2^3 = 10101000$$

Matching the 1s appearing in the binary expansion with the eight possible variable settings for three variables, we have the DNF expansion,

$$ABC \vee A\overline{B}C \vee \overline{A}BC$$

The corresponding *Mathematica* syntax for this Boolean expression is,

Or[And [p, q, r], And[p, Not[q], r], And[Not[p], q, r]]

In the text, we reduced the original complicated expression for proof by cases to $(A \vee B) \wedge C$. Translating this into *Mathematica* syntax,

And[Or[p, q], r]

Now in the same way as we have done before, set up the potential tautology between these two expressions to see whether they are logically equivalent,

**logicExpression2[p_, q_, r_] := Equivalent[Or[And [p, q, r],
      And[p, Not[q], r], And[Not[p], q, r]], And[Or[p, q], r]]**

Evaluating **TautologyQ[logicExpression2[p, q, r]]** returns **True**, and thus confirms that these two versions are indeed equivalent. This exercise also validates our rather more discursive discussion that occurred earlier.

Exercise 7.9.10. Explain how a joint probability table can be simplified if only one model is under consideration. Further generalize *modus ponens* in the process.

Solution to Exercise 7.9.10

In this Chapter, we have always had two models represented by $Z = T$ and $Z = F$. More specifically, the first model was where the logic function took on the functional assignment of T, and the second model was where the dual logic function switched these functional assignments. Therefore, when examining logic functions of two variables, an eight cell joint probability table involving A, B, and Z was introduced to account for the numerical assignments over both models.

If we were to consider just the *one* model where $Z = T$, then we could simplify the joint probability table to consist of just four cells. The first such 2×2 table is shown as Figure 7.8 as panel (1). If a model were to maintain a strict adherence to classical implication, then cell 3, $P(A\overline{B})$, must have a zero inserted as the numerical assignment. The other three cells might each have a numerical assignment of, say, $1/3$. Because there is only one model, the total probability of 1 can be partitioned out to just four cells.

The usual deductive result that B is TRUE if A is TRUE is reflected in,

$$P(B \,|\, A, A \to B) = 1$$

Figure 7.8: *Three 2×2 joint probability tables to illustrate variations on logical implication.*

As a first kind of generalization that deduction can not accomplish, we indicated that we might also want to look at the impact on A if B were TRUE with the result that $P(A \mid B, A \rightarrow B) = 1/2$.

But now generalize further by introducing the "almost an implication function" by permitting cell 3 to take on a small positive value, and letting the other three cells adjust accordingly as shown in panel (2). Here the model assigns $P(A\overline{B}) = 0.04$ instead of 0 as in the previous model. Now, if A is TRUE, B can no longer be TRUE from the deductive perspective of Classical Logic.

However, the inferential process is not thwarted as it provides this answer,

$$P(B \mid A, \mathcal{M}_2) = \frac{P(AB \mid \mathcal{M}_2)}{P(A \mid \mathcal{M}_2)} = \frac{0.32}{0.36} \approx 0.89$$

This probability as measure of belief says that while we cannot be certain that B is TRUE as we could before, there still exists more grounds for believing B TRUE than FALSE. The impact on A if B were TRUE remains the same at $1/2$ under this alternative model.

Finally, even this can change if a third model is introduced where once again the 0 for $P(A\overline{B})$ is replaced with the small value of 0.05 and the other three cells are given different numerical assignments as shown in panel (3). Now, although the ramifications for B are the same as before,

$$P(B \mid A, \mathcal{M}_3) = \frac{P(AB \mid \mathcal{M}_3)}{P(A \mid \mathcal{M}_3)} = \frac{0.40}{0.45} \approx 0.89$$

the impact on A changes from 0.50 to,

$$P(A \mid B, \mathcal{M}_3) = \frac{P(AB \mid \mathcal{M}_3)}{P(B \mid \mathcal{M}_3)} = \frac{0.40}{0.70} \approx 0.57$$

The information processor believes (possesses a state of knowledge) that A is now slightly more likely to be TRUE than FALSE under this model.

Exercise 7.9.11. Carry out the detailed Boolean operations needed to confirm the version of Bayes's Theorem appearing in section 7.7.

Solution to Exercise 7.9.11

Write out Bayes's Theorem for the problem as stated,

$$P(A \mid B, C, Z) = \frac{P(ABCZ)}{P(BCZ)}$$

The denominator is tackled first because it has one less variable. Begin with the CZ term and employ the shorter reformulation of the two implications,

$$C \wedge Z \;=\; C \wedge [\,(A \to B) \wedge (A \to C)\,]$$

$$A \to B \;=\; \overline{A} \vee B$$

$$A \to C \;=\; \overline{A} \vee C$$

$$(A \to B) \wedge (A \to C) \;=\; (\overline{A} \vee B) \wedge (\overline{A} \vee C)$$

$$(\overline{A} \vee B) \wedge (\overline{A} \vee C) \;=\; \overline{A} \vee (B \wedge C)$$

$$C \wedge Z \;=\; C \wedge [\,\overline{A} \vee (B \wedge C)\,]$$

$$\;=\; \overline{A}C \vee BCC$$

$$B \wedge C \wedge Z \;=\; \overline{A}BC \vee BBC$$

$$\;=\; \overline{A}BC \vee BC$$

The numerator simplifies even more because,

$$A \wedge B \wedge C \wedge Z \;=\; A \wedge [\,\overline{A}BC \vee BC\,]$$

$$\;=\; ABC$$

Thus, at this stage we have,

$$P(A \mid B, C, Z) = \frac{P(ABC)}{P(\overline{A}BC \vee BC)}$$

Refocus on the denominator and use the rule for distributing the probability operator across $\vee$.

$$P(\overline{A}BC \vee BC) = P(\overline{A}BC) + P(BC) - P(\overline{A}BCBC) = P(BC)$$

Now finish up by using the **Sum Rule** to expand the denominator,

$$P(BC) = P(ABC) + P(\overline{A}BC)$$

Bayes's Theorem appears in the form stated in section 7.7,

$$P(A \mid B, C, Z) = \frac{P(ABC \mid Z)}{P(ABC \mid Z) + P(\overline{A}BC \mid Z)}$$

Exercise 7.9.12. Using the model reflected in the joint probability table shown in Figure 7.5, determine the updated state of knowledge that a suspect might have the victim's car given that they have the victim's blood on their clothes.

Solution to Exercise 7.9.12

We return to making inferences within the context of our simplified crime scenario. We showed how an information processor updated its state of knowledge about a suspect being the murderer after being given the facts concerning the victim's car and blood. The IP had to assume some particular model of criminal activity before calculations could commence. The IP was not able to logically deduce anything about the crime, but could make inferences that flowed from a generalization of logic.

The question posed in this exercise is of exactly the same sort. It can be answered quantitatively through a probability calculation. For a suspect A who has the victim's blood on their person, and also given the model of criminality captured by the two implications, the probability that the suspect has the victim's car rises to nearly 50%.

This revised probability, $P(C \mid B, \mathcal{M}_k)$, based on the known fact B and some working hypothesis $\mathcal{M}_k$, can be calculated in the same manner as demonstrated for $P(A \mid B, C, \mathcal{M}_k)$ earlier. In short, these manipulations tell us that,

$$P(C \mid B, \mathcal{M}_k) = \frac{P(BC \mid \mathcal{M}_k)}{P(B \mid \mathcal{M}_k)}$$

What is $P(BC \mid \mathcal{M}_k)$? The joint probability table provides us with the numerical assignments made under model $\mathcal{M}_k$ and from the **Sum Rule** we know that,

$$P(BC) = P(ABC) + P(\overline{A}BC)$$

And plugging in the actual numbers from Figure 7.5,

$$P(BC) = 0.01 + 0.005 = 0.015$$

$P(B)$ is a marginal sum shown in the joint probability table as equal to 0.035.

Thus, the revised probability is seen to be,

$$
\begin{aligned}
P(C\,|\,B,\mathcal{M}_k) &= \frac{P(BC\,|\,\mathcal{M}_k)}{P(B\,|\,\mathcal{M}_k)} \\
&= \frac{0.015}{0.035} \\
&\approx 0.43
\end{aligned}
$$

From a relatively rare occurrence, it is now nearly an even bet, after establishing that the suspect's clothing is drenched in the victim's blood, that the suspect is in possession of the victim's car. Again, this statement is the consequence of assuming that one, and only one, particular model of criminal behavior is true. In reality, many, many models should be considered instead of just one until the day when enough data have accumulated to overwhelmingly support one, or a few, model(s).

Exercise 7.9.13. Discuss how these exercises lead one to the concept of averaging over models.

Solution to Exercise 7.9.13

We have looked at expressions of the form $P(B\,|\,A,Z)$ where the model Z was specified as some particular logic function, say, $A \to B$. Nothing prevents us from writing down $P(B\,|\,A)$ instead, where now Z has not been specified. Nonetheless, after expanding both AB and A, we have,

$$
P(B\,|\,A) = \frac{P(AB)}{P(A)} = \frac{P(ZAB) + P(\overline{Z}AB)}{P(ZAB) + P(\overline{Z}AB) + P(ZA\overline{B}) + P(\overline{Z}A\overline{B})}
$$

By requiring $Z = T$ or $Z = F$, we have essentially limited ourselves to two models. Because,

$$
P(A,B) = \sum_{k=1}^{\mathcal{M}} P(A,B,\mathcal{M}_k) = \sum_{k=1}^{\mathcal{M}} P(A\,|\,B,\mathcal{M}_k)\,P(B\,|\,\mathcal{M}_k)\,P(\mathcal{M}_k)
$$

we could then write,

$$
P(B\,|\,A) = \frac{\sum_{k=1}^{2} P(A\,|\,B,\mathcal{M}_k)\,P(B\,|\,\mathcal{M}_k)\,P(\mathcal{M}_k)}{\sum_B \sum_{k=1}^{2} P(A\,|\,B,\mathcal{M}_k)\,P(B\,|\,\mathcal{M}_k)\,P(\mathcal{M}_k)}
$$

This shows that the conditional probability of B is found by averaging with respect to the probabilities for *two* models.

Exercise 7.9.14. Conduct an inference about B by averaging over four logic functions, namely, $f_{13}(A, B)$, $f_{14}(A, B)$, and their respective dual functions.

Solution to Exercise 7.9.14

We are going to find $P(B \mid A)$ by averaging over four models. The joint probability tables have already been given for these logic functions and their duals in Figures 7.1 and 7.6. The dual function to f_{13} is f_4, $A \diamond B$, while the dual function to f_{14} is f_3, $A \star B$. These dual functions provide convenient models for $Z = F$ with their mirror image placement of assignments in the joint probability table.

Figure 7.9 is a sketch putting all four functions together into one 16 cell joint probability table. Note carefully that we have one 16 cell table with the models now part of the probability for joint statements written like $P(A, B, \mathcal{M}_k)$. The individual entries for the numerical assignments are either 1/8 or 0 instead of 1/4 and 0.

First, note some relevant marginal probabilities from this table. The overall sum of the entries is 1. The probabilities for the two arguments are,

$$P(A) = P(\overline{A}) = P(B) = P(\overline{B}) = 1/2$$

Figure 7.9: *16 cell joint probability table for $P(A, B, \mathcal{M}_k)$.*

The probabilities for the models are,

$$P(\mathcal{M}_1) = 3/8, P(\mathcal{M}_2) = 1/8, P(\mathcal{M}_3) = 3/8, P(\mathcal{M}_4) = 1/8$$

It is perhaps easier to calculate the probability $P(B \mid A)$ from first principles by examining the relevant cells of the newly enlarged joint probability table. The numerator in Bayes's Theorem is $P(AB) = \sum_{k=1}^{4} P(A, B, \mathcal{M}_k)$. After substituting,

$$P(AB) = 1/8 + 0 + 1/8 + 0 = 2/8$$

This sum, which will appear in the numerator, is the sum of the assignments in the four cells, cells 1, 5, 9, and 13.

The sum in the denominator will be these four cells, plus the four cells comprising $P(A\overline{B})$. These are cells 3, 7, 11, and 15. Together, the sum in these eight cells is,

$$P(A) = P(AB) + P(A\overline{B}) = [\, 1/8 + 0 + 1/8 + 0 \,] + [\, 0 + 1/8 + 1/8 + 0 \,] = 4/8$$

Bayes's Theorem then returns,

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{2/8}{4/8} = 1/2$$

We find the same answer by using a formula that explicitly shows the averaging over the models after the **Product Rule** is applied.

$$P(AB) = \sum_{k=1}^{4} P(A, B, \mathcal{M}_k) = \sum_{k=1}^{4} P(A \mid B, \mathcal{M}_k)\, P(B \mid \mathcal{M}_k)\, P(\mathcal{M}_k)$$

Taking $\mathcal{M}_1$ as an example, the three terms under the summation are,

$$P(A \mid B, \mathcal{M}_1) = \frac{P(A, B, \mathcal{M}_1)}{P(B, \mathcal{M}_1)} = \frac{1/8}{2/8}$$

$$P(B \mid \mathcal{M}_1) = \frac{P(B, \mathcal{M}_1)}{P(\mathcal{M}_1)} = \frac{2/8}{3/8}$$

$$P(\mathcal{M}_1) = 3/8$$

The numerator is then calculated as,

$$\sum_{k=1}^{4} P(A \mid B, \mathcal{M}_k)\, P(B \mid \mathcal{M}_k)\, P(\mathcal{M}_k) =$$

$$\left(\frac{1/8}{2/8} \times \frac{2/8}{3/8} \times 3/8 \right) + 0 + \left(\frac{1/8}{1/8} \times \frac{1/8}{3/8} \times 3/8 \right) + 0$$

$$= \quad \frac{2}{8}$$

The sum in the denominator will be these same four terms, plus the four terms over $\overline{B}$.

$$\sum_{k=1}^{4} P(A \mid \overline{B}, \mathcal{M}_k) \, P(\overline{B} \mid \mathcal{M}_k) \, P(\mathcal{M}_k) =$$

$$0 + \left(\frac{1/8}{1/8} \times \frac{1/8}{1/8} \times 1/8 \right) + \left(\frac{1/8}{2/8} \times \frac{2/8}{3/8} \times 3/8 \right) + 0$$

$$= \quad \frac{2}{8}$$

Thus,

$$P(B \mid A) \quad = \quad \frac{\sum_{k=1}^{4} P(A \mid B, \mathcal{M}_k) \, P(B \mid \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{B} \sum_{k=1}^{4} P(A \mid B, \mathcal{M}_k) \, P(B \mid \mathcal{M}_k) \, P(\mathcal{M}_k)}$$

$$= \quad \frac{2/8}{2/8 + 2/8}$$

$$= \quad 1/2$$

So, in the end, the IP was not able to update its state of knowledge about B after finding out about A. In its original state of knowledge, $P(B) = 1/2$. However, the new state of knowledge conditioned on A could not be revised upwards from the original state of knowledge. It remained at $P(B \mid A) = 1/2$. Missing information, reflected in the numerical assignments of the joint probability table, prevented the IP from making any progress.

Exercise 7.9.15. Use different joint probability tables to average over the joint probabilities of A and B.

Solution to Exercise 7.9.15

Suppose that we envision setting up joint probability tables that don't explicitly include the models as part of the joint statement. See Figure 7.10 at the top of the next page. What do the numerical assignments in each of the four cell joint probability tables for A and B now look like when conditioned on the four models? Figure 7.10 shows separate four cell joint probability tables with the assignments conditioned on assuming that one of the particular four models is true.

In other words, we don't want to divide up the probabilities over a 16 cell joint probability table by taking account of the four models, but simply want to generate smaller four cell joint probability tables for A and B as needed by assuming that some model is true. Now, as an obvious check, the sum of the numerical assignments in each of these four cell joint probability tables must sum to 1.

$$P(A,B \,|\, \mathcal{M}_1)$$

	A	**Ā**
B	**1/3** Cell 1	**1/3** Cell 2
B̄	**0** Cell 3	**1/3** Cell 4

$$P(A,B \,|\, \mathcal{M}_2)$$

	A	**Ā**
B	**0** Cell 1	**0** Cell 2
B̄	**1** Cell 3	**0** Cell 4

$$P(A,B \,|\, \mathcal{M}_3)$$

	A	**Ā**
B	**1/3** Cell 1	**0** Cell 2
B̄	**1/3** Cell 3	**1/3** Cell 4

$$P(A,B \,|\, \mathcal{M}_4)$$

	A	**Ā**
B	**0** Cell 1	**1** Cell 2
B̄	**0** Cell 3	**0** Cell 4

Figure 7.10: *Four cell joint probability tables for $P(A, B \,|\, \mathcal{M}_k)$ when conditioned on four models.*

We now show the derivation for $P(B \,|\, A)$ using these numerical assignments $P(A, B \,|\, \mathcal{M}_k)$ as conditioned on a particular model.

$$P(A, B) = \sum_{k=1}^{\mathcal{M}} P(A, B, \mathcal{M}_k)$$

$$= \sum_{k=1}^{\mathcal{M}} P(A, B \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)$$

$$P(B \,|\, A) = \frac{P(AB)}{P(A)}$$

$$= \frac{\sum_{k=1}^{\mathcal{M}} P(A, B \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}{P(A)}$$

$$= \frac{\sum_{k=1}^{\mathcal{M}} P(A, B \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(A \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}$$

$$P(B \,|\, A) = \frac{\sum_{k=1}^{\mathcal{M}} P(A, B \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} \left[\, P(A, B \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k) + P(A, \overline{B} \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k) \, \right]}$$

So we see in this last expression that the conditional probability for B is an average with respect to the probability for the models.

The numerator consists of the four terms making up $P(A, B \mid \mathcal{M}_k) \, P(\mathcal{M}_k)$. The assigned probabilities for the joint statements are plucked directly from cell 1 of the appropriate joint probability table as,

$$1/3, \; 0, \; 1/3, \; \text{and} \; 0$$

When multiplied by the probabilities for the models, the numerator becomes,

$$(1/3 \times 3/8) + (0 \times 1/8) + (1/3 \times 3/8) + (0 \times 1/8) = 2/8$$

The denominator consists of eight terms; these same four terms as just discussed, plus four more from $P(A, \overline{B} \mid \mathcal{M}_k) \, P(\mathcal{M}_k)$. The probabilities for the joint statements are plucked directly from cell 3 of the appropriate joint probability table as,

$$0, \; 1, \; 1/3, \; \text{and} \; 0$$

and when multiplied by the probabilities for the models these extra four terms are,

$$(0 \times 3/8) + (1 \times 1/8) + (1/3 \times 3/8) + (0 \times 1/8) = 2/8$$

All of this leads to the same answer that we found in the previous exercise,

$$P(B \mid A) = \frac{2/8}{2/8 + 2/8} = 1/2$$

Exercise 7.9.16. Is this the average of each Bayesian prediction?

Solution to Exercise 7.9.16

Extrapolating from the last two exercises, we might be led to conjecture that,

$$P(B \mid A) = \sum_{k=1}^{\mathcal{M}} P(B \mid A, \mathcal{M}_k) \, P(\mathcal{M}_k)$$

Checking with a numerical substitution, we find that,

$$P(B \mid A) = (1 \times 3/8) + (0 \times 1/8) + (1/2 \times 3/8) + (0 \times 1/8) = 9/16 \neq 1/2$$

What is wrong?

Start all over again from fundamental principles. Using the **Associativity** and **Commutativity** axioms, we have,

$$P(A, B, \mathcal{M}_k) \equiv P(B, \mathcal{M}_k, A)$$

Use the **Product Rule** on the right hand side to transform this into,

$$P(B, \mathcal{M}_k, A) = P(B \mid \mathcal{M}_k, A) \, P(\mathcal{M}_k \mid A) \, P(A)$$

Next, use the **Sum Rule** to find that,

$$P(A, B) = \sum_{k=1}^{\mathcal{M}} P(A, B, \mathcal{M}_k)$$

$$= \sum_{k=1}^{\mathcal{M}} P(B \mid \mathcal{M}_k, A)\, P(\mathcal{M}_k \mid A)\, P(A)$$

$P(A)$ is a constant that can be brought outside the summation,

$$P(A, B) = P(A) \sum_{k=1}^{\mathcal{M}} P(B \mid \mathcal{M}_k, A)\, P(\mathcal{M}_k \mid A)$$

Divide both sides by $P(A)$ and apply Bayes's Theorem to arrive at the desired result for $P(B \mid A)$.

$$P(A, B) = P(A) \sum_{k=1}^{\mathcal{M}} P(B \mid \mathcal{M}_k, A)\, P(\mathcal{M}_k \mid A)$$

$$\frac{P(A, B)}{P(A)} = \sum_{k=1}^{\mathcal{M}} P(B \mid \mathcal{M}_k, A)\, P(\mathcal{M}_k \mid A)$$

$$\frac{P(A, B)}{P(A)} = P(B \mid A)$$

$$P(B \mid A) = \sum_{k=1}^{\mathcal{M}} P(B \mid \mathcal{M}_k, A)\, P(\mathcal{M}_k \mid A)$$

$$P(B \mid A) = \sum_{k=1}^{\mathcal{M}} P(B \mid A, \mathcal{M}_k)\, P(\mathcal{M}_k \mid A)$$

Now when we conduct a numerical check, everything works out just fine. The averaging of each Bayesian prediction is carried out with respect to $P(\mathcal{M}_k \mid A)$,

$$P(\mathcal{M}_k \mid A) = \frac{P(A, \mathcal{M}_k)}{P(A)}$$

$$P(\mathcal{M}_1 \mid A) = \frac{1/8}{1/2} = 1/4$$

$$P(\mathcal{M}_2 \mid A) = \frac{1/8}{1/2} = 1/4$$

$$P(\mathcal{M}_3 \mid A) = \frac{2/8}{1/2} = 1/2$$

$$P(\mathcal{M}_4 \mid A) = \frac{0}{1/2} = 0$$

We must return to consulting the 16 cell joint probability table in Figure 7.9 to find $P(A, \mathcal{M}_k)$ because this is a probability concerning the joint statement of A and the model. The sum of the probabilities for all four models conditioned on A sums to 1 as it must,

$$\sum_{k=1}^{4} P(\mathcal{M}_k \mid A) = 1/4 + 1/4 + 1/2 + 0 = 1$$

and model $\mathcal{M}_4$ is eliminated when conditioned on the truth of A because A is impossible under that model. Thus,

$$P(B \mid A) = (1 \times 1/4) + (0 \times 1/4) + (1/2 \times 1/2) + (0 \times 0) = 1/2$$

Thankfully, this answer matches the result from the other two equivalent ways of approaching the problem.

Exercise 7.9.17. Use two axioms from Boolean Algebra to prove the analogous rearrangement of terms as depicted in the last exercise.

Solution to Exercise 7.9.17

This serves as a refresher exercise in the purely syntactic use of formal rules of manipulation. We want to show that variables in a Boolean Algebra expression can be rearranged to match the rearrangement shown in $P(A, B, \mathcal{M}_k) \equiv P(B, \mathcal{M}_k, A)$. Using only the **Associativity** and **Commutativity** axioms, we will prove an easy theorem.

Wolfram has clarified our conceptual understanding of theorem proving. He tells us to view theorem proving as simply constructing two columns, separated by the symbol $\Rightarrow$, showing permissible transformations on a starting string. Each permissible transformation produces a new string. The assembled collection of allowable transformations eventually leads to some final string. The final string produced is the desired theorem.

The permissible transformations consist of any starting axioms, together with any lemmas or theorems proved previously. Here, in this simple theorem, only the two starting axioms just mentioned are used to transform strings. The theorem is stated as,

$$(x \circ y) \circ z = (y \circ z) \circ x$$

Following Wolfram's prescription, the proof of this theorem is then,

$$(x \circ y) \circ z \quad \Rightarrow \quad (y \circ x) \circ z$$

$$(y \circ x) \circ z \quad \Rightarrow \quad y \circ (x \circ z)$$

$$y \circ (x \circ z) \quad \Rightarrow \quad y \circ (z \circ x)$$

$$y \circ (z \circ x) \quad \Rightarrow \quad (y \circ z) \circ x$$

Suppose we give the following values to the variables from some carrier set **B**,

$$x = a$$

$$y = b$$

$$z = T$$

Then, from the theorem just derived, $(a \circ b) \circ T$ must be the same as $(b \circ T) \circ a$.

Furthermore, if the formal manipulation rules of probability theory are analogous to Boolean Algebra, then $P(A, B, \mathcal{M}_k)$ must equal $P(B, \mathcal{M}_k, A)$.

Chapter 8

Deterministic Cellular Automata

8.1   Introduction

In the last Chapter, we showed how probability could generalize Classical Logic. We voluntarily restricted ourselves for the most part to functions with two arguments. In what amounts to the same thing, we were limited to probabilities involving three statements by taking the functional assignment to these two arguments into account.

We found that the few syllogisms involving more complicated logic functions caused no particular difficulty. Would we anticipate any problems in extending these lessons about generalization to Boolean functions with three or more arguments?

Chapter Three introduced the idea of cellular automata, and more especially Wolfram's 256 elementary CA. These CA were seen to be composed from Boolean functions defined on three arguments with the functional assignment the same as the color of the updated cell. Thus, we are now thinking about probabilities involving four statements.

CA were visually motivated by thinking of a window extending across three cells moving over the line of cells at the previous time step in order to assign a new color to a cell at the current time step. This operation was repeated in a machine–like fashion horizontally across a one–dimensional string of cells, and then vertically down a discrete time dimension.

This visual interpretation of a CA was augmented by coloring the cells black or white to correspond to assignments of T or F. Since the machinery behind these elementary CA is analogous to logic functions defined with three arguments, they should be amenable to a similar treatment by the formal manipulation rules of probability. Once again, Bayes's Theorem is relied upon as the most powerful way to apply these formal rules.

The ultimate goal is to conduct the same kind of attack on the elementary CA as was conducted on Classical Logic. Eventually, we will want to focus in on the particular CA implemented as Rule 110 simply because of its unbelievable capability as a Universal Turing Machine. First though, we develop in some detail a familiarity with the notation to be used in upcoming Chapters on CA and probability.

This Chapter restricts itself to deterministic CA. That is, even though we couch everything in probabilistic terms, the models duplicate the output from the direct application of the deterministic rules as first laid out in Chapter Three. That is, the probability of coloring cell B_{N+1} black or white must be 1 or 0. The obvious next step in generalizing these deterministic CA is taken up in the next Chapter.

8.2   Enforcing Determinism

To begin, and as a refresher, we will continue the story begun in the last Chapter. That is, we will start off by considering probabilities involving the two arguments of some logic function. But we will phrase the questions in a slightly different manner than before.

Previously, we placed a statement Z to the *right* of the conditioned upon symbol as in $P(A \,|\, B, Z)$. The situation where Z stood for the implication operator,

$$Z = T \equiv A \rightarrow B \equiv f_{13}(A, B)$$

was the canonical example. More generally, we will eventually want to use $\mathcal{M}_k$ in place of Z to indicate that some model has made numerical assignments instead of assuming some particular logic function. Then, we would write it as $P(A \,|\, B, \mathcal{M}_k)$.

But, as we learned, strict determinism existed in only one direction for this *deductive* system. B was known with certainty only if A were true. Curiously though, in the reverse direction, A was essentially undecidable even if B were known while still assuming that the logic function $A \rightarrow B$ was the model. However, an *inferential* system could answer questions about A given knowledge about B, and might even approach certainty if probabilities for A and B were assigned specific values under some model.

Nothing in what we have just said is changed if we remember that statements can be moved around at will within probability expressions, together with adopting a flexible attitude toward the abstract notation. We might just as well have said that the functional assignment was a statement C. Thus, expressions like $P(A \,|\, B, C)$ have already appeared, but $P(C \,|\, A, B)$ is acceptable as well.

It's easy to discern the pattern that must exist in order to enforce determinism. Suppose we are trying to determine the degree of belief about statement C given that A and B are known. Bayes's Theorem would be written as,

$$P(C \,|\, A, B) = \frac{P(ABC)}{P(AB)} = \frac{P(ABC)}{P(ABC) + P(AB\overline{C})}$$

To enforce determinism, either $P(AB\overline{C})$ must be 0 resulting in $P(C \,|\, A, B) = 1$ or $P(ABC)$ must be 0 resulting in $P(C \,|\, A, B) = 0 \equiv P(\overline{C} \,|\, A, B) = 1$.

Extending these thoughts to the elementary CA, we now have the situation where the functional assignment is labeled as D, and the logic function has three arguments, A, B, and C. D is the color of the updated cell at time step $N + 1$, while A, B, and C are the colors of the neighboring cells at the previous time step.

Likewise, suppose we are trying to determine the degree of belief about statement D given that A, B, and C are known. In other words, the IP wants to determine its degree of belief about the updated cell's color given that it knows the color of the neighboring cells. Since this is a deterministic system, there is no uncertainty about the updated cell's color.

Bayes's Theorem would be written as,

$$P(D \,|\, A, B, C) = \frac{P(ABCD)}{P(ABCD) + P(ABC\overline{D})}$$

To enforce determinism, either $P(ABC\overline{D})$ must be 0 resulting in $P(D \,|\, A, B, C) = 1$ (the updated cell is black), or $P(ABCD)$ must be 0 resulting in $P(D \,|\, A, B, C) = 0$ (the updated cell is white).

8.2.1 A joint probability table enforcing determinism

We've employed joint probability tables for logic functions with two arguments before. Figure 8.1 shown below is, at least, a conceivable joint probability table since legitimate numerical assignments are made to all eight cells.

Remember that this particular numerical assignment takes place because of the information resident in some model. Since we are emphasizing the manipulation

Figure 8.1: *Joint probability table inspired by a logic function and its dual.*

rules here in Volume I rather than any numerical assignment algorithm, take these numbers as if they were received from on High. That is, we are not going into any of the details of the algorithm which would produce these numbers. Suffice it to say that the OR logic function, and its dual function NOR, are the inspiration for the numerical assignments made by the model.

If so desired, one could look at this as two four cell joint probability tables. From this perspective, each table has numerical assignments conditioned on C or $\overline{C}$. Or, C and $\overline{C}$ correspond to two different models. Thus, instead of viewing the joint probability table as one eight cell table, it is thought of as two four cell tables, one for each model. If we do that, then the numerical assignments must be re–adjusted so that the sums in each four cell table sum to 1.

There are four zeroes in the joint probability table, one zero located among the four cells on the left hand side and three zeroes strategically placed among the right hand cells. The model forced the left hand side to mimic the OR logic function, and the right hand side to mimic the function dual to OR, the NOR logic function.

These strategically placed zeroes enforce determinism. Our previous example asked about the status of C given knowledge of both A and B. C is known with certainty because of the zero in cell 5.

$$P(C \,|\, A, B) = \frac{P(ABC)}{P(ABC) + P(AB\overline{C})} = 1$$

This example was designed to get us in the mood for thinking about deterministic CA. These CA are equivalent to functional assignments ensuing from logic functions with three arguments. This means that we add another variable to our current discussion so that we are effectively dealing with probabilities about four statements and 16 cell joint probability tables.

8.3   Confused About Joint Probability Tables?

To be honest, I haven't been very clear about these joint probability tables. Let's try to disambiguate the confusion surrounding the number of cells in a joint probability table.

Up till now, we have tended to write the probability solely for joint statements. Thus, $P(BAZ)$ represented a state of knowledge about three binary statements. A and B referred to the two arguments of a logic function, while Z was a statement about which logic function was in play.

Any numerical assignments were shown in an eight cell joint probability table since there were just two values for A, B, and Z. We wrote expressions like $P(BAZ)$ in Bayes's Theorem without any conditionalization notation explicitly present. But we labored trying to explain that the two values for Z, $Z = T$ and $Z = F$, could just as well be thought of as two models making numerical assignments.

Thus, Bayes's Theorem used only joint probabilities on the right hand side, as in the canonical example,

$$P(B \mid A, A \to B) = \frac{P(BAZ)}{P(BAZ) + P(\overline{B}AZ)}$$

Then, in this Chapter, we wrote more general expressions like $P(A \mid B, C)$ where A, B, and C are all statements, but statement C was thought to represent a model. For this situation, we considered the option of reducing the original eight cell joint probability table into two four cell joint probability tables, where now each four cell table was conditioned on one of two models. The better notation was now $P(A, B \mid \mathcal{M}_1)$ or $P(A, B \mid \mathcal{M}_2)$ for each four cell table.

Follow this same pattern as we increase the number of statements. For four statements, $P(A, B, C, Z)$, we could construct one 16 cell joint probability table, or two 8 cell tables, $P(A, B, C \mid \mathcal{M}_1)$ and $P(A, B, C \mid \mathcal{M}_2)$. For five statements, $P(A, B, C, D, Z)$, we could construct one 32 cell joint probability table, or two 16 cell tables, $P(A, B, C, D \mid \mathcal{M}_1)$ and $P(A, B, C, D \mid \mathcal{M}_2)$.

We can stop here for our purposes in dealing with the 256 elementary cellular automata. If we think of statement D as the functional assignment based on the three arguments A, B, and C, we can get away with 16 cell joint probability tables. This is true only if we realize that we are also conditioning on just one model $\mathcal{M}_k$ to construct this 16 cell table.

In other words, we will be using Bayes's Theorem with expressions like,

$$P(D \mid A, B, C, \mathcal{M}_k)$$

on the left hand side involving four statements and a given model, just as we could use a four cell table when only two statements and a given model, $P(A \mid B, \mathcal{M}_1)$, were involved.

8.4   Inferences About Four Statements

Now that we have cleared away all of this undergrowth, we can directly attack the deterministic CA. Of course, nothing in the formal rules for probability prevents us from considering an extension to four statements. As always, these statements are referred to generically as A, B, C, and D.

We will, for the moment, focus on the idea that D is the statement about the functional assignment from some logic function with three arguments A, B, and C. Again, we are restricting ourselves to arguments and functional assignments of T and F, or colors of black and white.

The joint probability table gets bumped up to 16 cells if the assignments are made under one model. However, the symmetry we recognized as necessary for a

deterministic system remains the same. There will be eight zeroes scattered about the joint probability table with the positive values under D balanced off by zeroes in the corresponding cells under $\overline{D}$, and *vice versa*.

Thus, we don't even have to explicitly show all of the detail in this 16 cell joint probability table for the upcoming numerical example. It is enough to exploit the symmetry that must exist for a deterministic system, together with Bayes's Theorem. We have noticed in previous examples that we could pick out a two variable logic function and its dual as a convenient model for numerical exercises. We do the same thing here.

Select some three variable logic function for the cells under D and then determine its dual for the cells under $\overline{D}$. In this manner, we can avoid having to find some four variable logic function, although it would not be particularly difficult to do so. Basically, we'd like to retain the terminology whereby we can keep referring to Wolfram's rule numbers for elementary CA.

Because we have used it previously as an example in Chapter Three, take Rule 192 as the inspiration for a three variable logic function. Make it assign positive values and zeroes for the cells under D. For our purpose here, we don't even need to know what the dual function is, just balance off the cells under D with appropriate values and zeroes for the cells under $\overline{D}$.

Having accomplished this, Bayes's Theorem, within its general probabilistic mandate, will do its job and recreate the deterministic output. Suppose we were interested in the probability of the functional assignment taking on the value of T when A takes on the value T, B takes on the value F, and C takes on the value T.

$$P(D\,|\,A,\overline{B},C) = \frac{P(A,\overline{B},C,D)}{P(A,\overline{B},C,D) + P(A,\overline{B},C,\overline{D})}$$

As mentioned above, the joint probability table under this model will consist of 16 cells. There are eight cells for the various combinations under D, and the remaining eight cells for the various combinations under $\overline{D}$. Once again, use the DNF for Rule 192 to determine just where in the joint probability table for D some positive probability may be placed or, alternatively, where an F, and thus a 0 must be placed.

Now if we go back to Chapter Three to refresh our memory as to the DNF expansion of Rule 192, we find that it was simply $ABC \vee AB\overline{C}$. The DNF pinpoints the two cells of the joint probability table for D where a probability doesn't have to be 0. Keeping in mind the fact that the arguments to the function are now A, B, and C, and $D = T$, those two cells are $P(ABCD)$ and $P(AB\overline{C}D)$. The remaining six cells under D must contain 0s.

Furthermore, the appropriate six cells under $\overline{D}$ will have values different than 0, while the remaining two cells under $\overline{D}$ will contain 0s. To find the denominator in Bayes's Theorem, we are interested in cell 3, $P(A\overline{B}CD)$, and cell 11, $P(A\overline{B}C\overline{D})$.

Cell 3 has a value of 0 since only cell 1, $P(ABCD)$, and cell 5, $P(AB\overline{C}D)$, contained non–zero assignments. Say cell 11 has some value q since it is the symmetrical counterpart to cell 3 which contained 0. Substituting these values into Bayes's Theorem we have,

$$P(D \mid A, \overline{B}, C) = \frac{0}{0 + q} = 0$$

The functional assignment of F must be made for D when the variable settings of the arguments are $A = T$, $B = F$, and $C = T$ together with the models inspired by Rule 192 and its dual. This model is driving the numerical assignments that are placed into the cells of the joint probability table appearing under D and $\overline{D}$.

It was easily proved in Chapter Three that the DNF expansion for Rule 192 reduced to AB. This the same as the logic function `AND`. So, it is clear that if we have the statements to the right of the conditioned upon symbol to the effect that $A = T$ and $B = F$, D cannot have a functional value of T.

8.5 Switch the Notation to Cellular Automata

We now shift our attention from logic functions involving three arguments together with probabilities of four joint statements to elementary cellular automata. Such a translation takes place seamlessly as it really only demands a straightforward notational change.

The variables A, B, and C correspond to the three neighboring cells A_N, B_N, and C_N at the Nth time step evolution. D becomes the updated cell B_{N+1} at the next time step. The variable settings, as well as the functional assignment with the variables as arguments, are no longer T and F, but become the cell colors, black and white.

In all other respects, the analysis remains the same. Thus, given that some CA rule will serve as inspiration for the numerical assignments in the joint probability table, Bayes's Theorem is written as,

$$P(B_{N+1} \mid A_N, B_N, C_N) = \frac{P(B_{N+1}, A_N, B_N, C_N)}{P(B_{N+1}, A_N, B_N, C_N) + P(\overline{B}_{N+1}, A_N, B_N, C_N)}$$

Since the current context involves CA, we will use language that asks for the probability that the updated cell is white, given that the cell was white at the previous time step and its two neighbor cells were black. Bayes's Theorem for this example is then written as,

$$P(B_{N+1} = w \mid A_N = b, B_N = w, C_N = b) =$$

$$\frac{P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N)}{P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N) + P(B_{N+1}, A_N, \overline{B}_N, C_N)}$$

Additionally, pick any one of Wolfram's 256 elementary rules to dictate the numerical assignments in the cells of the joint probability table. Say, just off the top of our heads, we pick Rule 221.

Now given our discussion of deterministic CA, we know that the probability we seek must be either 1 or 0, that is, the updated cell must be colored black (b) or white (w). This deduction is duplicated by the Bayes's Theorem template where we have substituted the notation for white and black cells. The IP knows that the template will take the form of either,

$$P(w \mid b, w, b) = \frac{q}{q+0} = 1$$

or,

$$P(w \mid b, w, b) = \frac{0}{0+q} = 0$$

The probability for a white cell is seen to depend on the placement of the joint probabilities and, more significantly, on the placement of the 0s in the joint probability table as dictated by the DNF expansion of Rule 221. You can look at Exercise 8.6.8 for all of the details that complete the solution to this problem.

In the end, it is better to conceptualize one model for the CA as based on one 16 cell joint probability table, rather than as two models based on two eight cell tables. Thus, Bayes's Theorem would be written as,

$$P(B_{N+1} \mid A_N, B_N, C_N, \mathcal{M}_k) =$$

$$\frac{P(B_{N+1}, A_N, B_N, C_N \mid \mathcal{M}_k)}{P(B_{N+1}, A_N, B_N, C_N \mid \mathcal{M}_k) + P(\overline{B}_{N+1}, A_N, B_N, C_N \mid \mathcal{M}_k)}$$

to indicate that the numerical answer is clearly dependent on this one k^{th} model, the $\mathcal{M}_k$ dictating the numerical assignments to the joint probability table.

Any of the 256 logic functions with three arguments can serve as a heuristic to fill in the first half of the joint probability table, while the dual function can pinpoint the placement of assignments in the second half. As always, these assignments to the B_{N+1} and $\overline{B}_{N+1}$ halves of the table will mimic, from a probabilistic perspective, any deterministic CA system.

8.6 Solved Exercises for Chapter Eight

Exercise 8.6.1. Begin by supposing that the NAND logical operator is used as inspiration for a model. Given that A and B are both TRUE, what is the status of C?

Solution to Exercise 8.6.1

This problem is slightly different than previous ones considered in Chapter Seven in that we are really inquiring about the state of knowledge concerning two models. Statement C replaces statement Z, but Bayes's Theorem is easily written down.

The formal manipulation rules tell us that,

$$P(C \mid A, B) = \frac{P(ABC)}{P(AB)}$$

Since $C = T$, substitute an equivalent shorter version for the DNF of the NAND logic function, $\overline{(A \wedge B)} \equiv \overline{A} \vee \overline{B}$,

$$P(C \mid A, B) = \frac{P(A \wedge [\, B \wedge [\, \overline{A} \vee \overline{B} \,]\,])}{P(AB)}$$

Carry out the Boolean operations in the numerator,

$$
\begin{aligned}
P(A \wedge [\, B \wedge [\, \overline{A} \vee \overline{B} \,]\,]) &= P(A \wedge [\, B\overline{A} \vee B\overline{B} \,]) \\
&= P(A \wedge [\, \overline{A}B \,]) \\
&= P(A\overline{A}B) \\
&= P(F) \\
&= 0
\end{aligned}
$$

Since $P(C \mid A, B) = 0$, the first model is ruled out. This is easily understood. If the two arguments to a logic function are asserted unambiguously as $A = T$ and $B = T$ (they are placed to the right of the conditioned upon symbol), then it is impossible by definition that NAND can be that logic function. Furthermore, in our universe of discourse, there exist only two models. If $P(C \mid A, B) = 0$, then $P(\overline{C} \mid A, B) = 1$. It must be true that the AND logic function is the correct model.

Exercise 8.6.2. Review the full DNF expansion for `NAND`. Determine whether there is a function among the orthonormal basis functions where the coefficient to the basis function is **FALSE**, but the arguments are both **TRUE**.

Solution to Exercise 8.6.2

Boole's Expansion Theorem for `NAND`, that is, the expansion showing the coefficients for the orthonormal basis functions, yields the following,

$$f_{12}(A, B) = f(T, T) \, AB \vee f(T, F) \, A\overline{B} \vee f(F, T) \, \overline{A}B \vee f(F, F) \, \overline{A}\,\overline{B}$$

After inserting the correct coefficients for the `NAND` operator, which are, in order, F, T, T, and T, the full DNF expansion is $(A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$, "$B$ is FALSE, or A is FALSE, or both are FALSE."

These three terms represent the three variable settings where the functional assignment of TRUE is made by `NAND`. The first term in the DNF expansion, AB, has a coefficient of FALSE.

Exercise 8.6.3. Think up a legitimate joint probability table for the situation in the previous exercises and comment on the placement of zeroes.

Solution to Exercise 8.6.3

Figure 8.2 shows one legitimate assignment of numerical values to all eight cells of a joint probability table for three variables. The four left hand cells are motivated by the `NAND` logical function, and the four right hand cells by the function dual to `NAND`, the `AND` logic function.

Figure 8.2: *Possible joint probability table for three statements inspired by the* `NAND` *operator and its dual* `AND`.

There are four zeroes scattered about the eight cells of the joint probability table, one zero on the left hand side, balanced off by three zeroes on the right hand side. Some marginal probabilities are $P(A) = 1/2$, $P(\overline{A}) = 1/2$, $P(B) = 7/12$, $P(\overline{B}) = 5/12$, and $P(C) = 2/3$, $P(\overline{C}) = 1/3$.

If A and B are both TRUE, then the functional assignment of T is not possible under the NAND operator. The DNF expansion for the NAND operator told us this.

On the other hand, if A and B are both TRUE, then the functional assignment of T must be made under the AND operator. The other possibilities for AND have a functional assignment of F. That is why we see zeroes in these three cells on the right hand side. It is balanced off by the one zero in the first cell on the NAND side where a functional assignment of F is made by that operator to $A = T$ and $B = T$.

Exercise 8.6.4. Reduce the full DNF expansion for the NAND operator to the shorter expression as used in Exercise 8.6.1.

Solution to Exercise 8.6.4

By **De Morgan's axiom**,

$$\overline{A} \vee \overline{B} = \overline{(A \wedge B)} = \overline{AB}$$

But $\overline{AB}$ is the definition for NAND, and expressed in unambiguous *Mathematica* notation, is written as **Not[And[p, q]]**. What are, in fact, the three cases that constitute $\overline{AB}$? Or, what are the three cases where A and B are not both TRUE? Here are those three cases, the DNF expansion for NAND,

$$(A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$$

A *Mathematica* program can prove the logical equivalency between the full DNF expression for the NAND logic function and the shorter version.

Exercise 8.6.5. Using the joint probability table from Exercise 8.6.3, what is the state of knowledge that B is FALSE given that A is FALSE?

Solution to Exercise 8.6.5

The easiest way to solve this is through,

$$P(\overline{B}\,|\,\overline{A}) = \frac{P(\overline{A}\,\overline{B})}{P(\overline{A})} = \frac{P(C\overline{A}\,\overline{B}) + P(\overline{C}\,\overline{A}\,\overline{B})}{P(C\overline{A}\,\overline{B}) + P(\overline{C}\,\overline{A}\,\overline{B}) + P(C\overline{A}B) + P(\overline{C}\,\overline{A}B)}$$

After substituting the numerical assignments from the joint probability table,

$$P(\overline{B}\,|\,\overline{A}) = \frac{1/4 + 0}{1/4 + 0 + 1/4 + 0} = 1/2$$

These numerical assignments were plucked from the following cells of the joint probability table,

$$P(\overline{B}\,|\,\overline{A}) = \frac{\text{Cell } 4 + \text{Cell } 8}{\text{Cell } 4 + \text{Cell } 8 + \text{Cell } 2 + \text{Cell } 6}$$

Exercise 8.6.6. Verify the answer in the previous exercise, but now look at it from the model averaging perspective.

Solution to Exercise 8.6.6

It is imperative that the same answer must be calculated from the model averaging perspective. Exercise 7.9.16 derived a formula to highlight the averaging of Bayesian predictions. Applying that formula here, we have,

$$P(\overline{B}\,|\,\overline{A}) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}\,|\,\overline{A}, \mathcal{M}_k)\, P(\mathcal{M}_k\,|\,\overline{A})$$

After applying Bayes's Theorem, the first term in the summation looks like,

$$P(\overline{B}\,|\,\overline{A}, \mathcal{M}_k) = \frac{P(\overline{A}\,\overline{B}\,|\,\mathcal{M}_k)}{P(\overline{A}\,|\,\mathcal{M}_k)}$$

Likewise, the second term looks like,

$$P(\mathcal{M}_k\,|\,\overline{A}) = \frac{P(\overline{A}\,|\,\mathcal{M}_k)\, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\overline{A}\,|\,\mathcal{M}_k)\, P(\mathcal{M}_k)}$$

Since there are only $\mathcal{M} = 2$ models,

$$P(\overline{B}\,|\,\overline{A}) = \left[\frac{P(\overline{A}\,\overline{B}\,|\,\mathcal{M}_1)}{P(\overline{A}\,|\,\mathcal{M}_1)} \times \frac{P(\overline{A}\,|\,\mathcal{M}_1)\, P(\mathcal{M}_1)}{P(\overline{A}\,|\,\mathcal{M}_1)\, P(\mathcal{M}_1) + P(\overline{A}\,|\,\mathcal{M}_2)\, P(\mathcal{M}_2)}\right] +$$

$$\left[\frac{P(\overline{A}\,\overline{B}\,|\,\mathcal{M}_2)}{P(\overline{A}\,|\,\mathcal{M}_2)} \times \frac{P(\overline{A}\,|\,\mathcal{M}_2)\, P(\mathcal{M}_2)}{P(\overline{A}\,|\,\mathcal{M}_1)\, P(\mathcal{M}_1) + P(\overline{A}\,|\,\mathcal{M}_2)\, P(\mathcal{M}_2)}\right]$$

Refer once again to the joint probability table in Figure 8.2 and Exercise 8.6.3 to find the numerical assignments under each model. The statements C and $\overline{C}$ are now thought of as the two statements $\mathcal{M}_1$ and $\mathcal{M}_2$.

$$\begin{aligned} P(\overline{B}\,|\,\overline{A}) &= \left[\frac{1/4}{1/2} \times \frac{(1/2 \times 2/3)}{(1/2 \times 2/3) + (0 \times 1/3)}\right] + 0 \\ &= 1/2 \end{aligned}$$

Exercise 8.6.7. Provide a summary discussion for the placement of the zeroes in the joint probability tables for all sixteen logic functions.

Solution to Exercise 8.6.7

There is a satisfying combinatorial explanation relating where the zeroes appear in the joint probability tables for the sixteen logic functions. Consider first the four logic functions where the DNF expansion indicates that three out of the four possible variable settings take on the functional assignment of T. These four logic functions are, in fact, the ones often used as examples.

1. NAND $\uparrow$

2. IMPLIES $\leftarrow$

3. IMPLIES $\rightarrow$

4. OR $\vee$

If three out of the four variable settings take on the functional assignment of T for these logic functions, then one of the four variable settings takes on the functional assignment of F. Consequently, one 0 will be placed in the appropriate cell of the joint probability table. The placement of the 0 is systematically rotated from cell 1 to cell 2 to cell 3 to cell 4 following the order given in the list above.

Combinatorially, how many ways are there to place one item, the 0, into four cells? There must be $\binom{4}{1} = 4$ ways to do this. And, in fact, these four ways are simply where the 0 is in cell 1 for NAND, in cell 2 for IMPLIES $(B \rightarrow A)$, in cell 3 for IMPLIES $(A \rightarrow B)$, and finally in cell 4 for OR.

Retain this mode of thinking when contemplating where the 0s would go for the other functions. Thus, it is easy to see that the number of logic functions where the DNF expansion consists of two terms must be $\binom{4}{2} = 6$ with 6 different ways to locate the two 0s in the four cells of the joint probability table.

1. NOT B $\dashv$

2. NOT A $\vdash$

3. XOR $\oplus$

4. EQUAL $\leftrightarrow$

5. A $\lhd$

6. B $\rhd$

For example, logic function $f_{11}(A, B)$ was called A and also represented notationally as $A \lhd B$. The DNF expansion for this function was $AB \vee A\overline{B}$ where A assumes

the functional assignment of T at each of these two variable settings. Therefore, A has the functional assignment of F, and a numerical assignment of 0 for $P(\overline{A}B)$ and $P(\overline{A}\,\overline{B})$. These probabilities are located in cells 2 and 4 of the joint probability table.

Trying to visualize all six patterns of two 0s, we see that the two 0s could be in the top row, the bottom row, the right column, the left column, or diagonally in two ways. This exhausts all the possibilities for placing two 0s in four cells. In the list above, the two 0s appear respectively in (1) cells 1 and 2, (2) cells 1 and 3, (3) cells 1 and 4, (4) cells 2 and 3, (5) cells 2 and 4 and, finally, (6) cells 3 and 4.

Rounding things out, there are $\binom{4}{3} = 4$ functions where the DNF expansion has only one term. Consequently, there are three Fs and three 0s to be placed into the joint probability table. AND is an example of such a function that has the functional assignment of T only when $A = T$ and $B = T$.

1. NOR $\downarrow$

2. DIFFERENCE $\diamond$

3. DIFFERENCE $\star$

4. AND $\wedge$

In the list above, the three 0s appear respectively in (1) cells 1, 2, and 3, (2) cells 1, 2, and 4, and (3) cells 1, 3, and 4, and, finally, (4) cells 2, 3, and 4.

We have checked off 14 of the 16 possible functions, leaving only two functions unaccounted for. One is the $\binom{4}{4} = 1$ way to place all four 0s into the four cells, and the second is the $\binom{4}{0} = 1$ way to place no 0s into the four cells. These last two functions are obviously $f_1(A, B)$, the $\bot$ operator, and $f_{16}(A, B)$, the $\top$ operator.

Exercise 8.6.8. Complete the development of the notation for Bayes's Theorem to cellular automata as begun in section 8.5.

Solution to Exercise 8.6.8

First, decode Wolfram's numbering system for his elementary cellular automata into a binary number. For example, Rule 221 would be,

$$
\begin{aligned}
221 \;=\; & (1 \times 2^7) + (1 \times 2^6) + (0 \times 2^5) + (1 \times 2^4) + \\
& (1 \times 2^3) + (1 \times 2^2) + (0 \times 2^1) + (1 \times 2^0) \\
=\; & 11011101
\end{aligned}
$$

Next, this binary number tells us that the DNF expansion for this logic function is,

$$ABC \vee AB\overline{C} \vee A\overline{B}\,\overline{C} \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}$$

We now use this DNF expansion to locate those 16 cells in the joint probability table that have some probability and those that have 0s. For example, the first cell, $P(B_{N+1}A_N B_N C_N)$, might have some probability q. The corresponding cell, cell 9, $P(\overline{B}_{N+1}A_N B_N C_N)$, will contain a 0. We won't explicitly show the full joint probability table, but figure out what probabilities we'll require from it after writing out Bayes's Theorem.

Section 8.5 presented Bayes's Theorem in terms of statements being observed as white or black cells. Before we get to that point, let's write out Bayes's Theorem in the more familiar style of generic statements, and then present the mapping that translates between the two representations. $\mathcal{M}_k$ indicates that the numerical assignments in the 16 cell joint probability table are based on Rule 221 and its dual,

$$P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, C_N, \mathcal{M}_k) =$$

$$\frac{P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N \,|\, \mathcal{M}_k)}{P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N \,|\, \mathcal{M}_k) + P(B_{N+1}, A_N, \overline{B}_N, C_N \,|\, \mathcal{M}_k)}$$

Here is the mapping between the notations:

$$\overline{B}_{N+1} \Rightarrow (B_{N+1} = F) \Rightarrow w$$

$$A_N \Rightarrow (A_N = T) \Rightarrow b$$

$$\overline{B}_N \Rightarrow (B_N = F) \Rightarrow w$$

$$C_N \Rightarrow (C_N = T) \Rightarrow b$$

so that we might rewrite Bayes's Theorem, as indicated earlier in the Chapter, as,

$$P(w \,|\, b, w, b, \mathcal{M}_k) = \frac{P(w, b, w, b \,|\, \mathcal{M}_k)}{P(w, b, w, b \,|\, \mathcal{M}_k) + P(b, b, w, b \,|\, \mathcal{M}_k)}$$

We didn't see $A\overline{B}C$ in the full DNF expansion, therefore the appropriate cell in the joint probability table under B_{N+1} will have a probability of 0 assigned. The corresponding cell in the joint probability table referring to the first term in the denominator, which is also the numerator, has an assigned probability q. Plugging these observations into Bayes's Theorem, it is certain that the updated cell is white.

$$P(w \,|\, b, w, b, \mathcal{M}_k) = \frac{q}{q+0} = 1$$

Exercise 8.6.9. Employing the same kind of combinatorial argument as in Exercise 8.6.7, figure out how many 0s would appear in the joint probability tables for Wolfram's elementary CA.

Solution to Exercise 8.6.9

The combinatorial formula useful here is,

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

In the previous exercise, we looked at all the logic functions of two variables where,

$$\sum_{k=0}^{4} \binom{4}{k} = 2^4 = 16$$

For the elementary CA, $n = 8$ so that

$$\sum_{k=0}^{8} \binom{8}{k} = 2^8 = 256$$

and $\binom{8}{k}$ will tell us how many functions have $8 - k$ terms in the DNF expansion. Also, it will tell us in how many different ways the k 0s can be allocated to the eight cells of the 16 cell joint probability tables under B_{N+1}.

Exercise 8.6.10. Discuss some of the classes of logic functions with three arguments centering around the notion of the number of 0s in the joint probability table.

Solution to Exercise 8.6.10

Again, it is easiest to first consider the eight functions $\binom{8}{1} = 8$ where the DNF has seven terms where the functional assignment is T. There is only one term where the assignment is an F. This one 0 can be rotated, in turn, from cell 1 through cell 8 in eight ways.

Consider the function among these eight that has the 0 in cell 1. Choosing this cell for the placement of the one 0 implies that the probability for the joint statement that the updated cell is black and the three relevant cells at the previous time step were also all black, that is $P(B_{N+1}A_N B_N C_N)$, must be 0.

What rule would this be? Well, if the DNF expansion has a F for cell 1 and a T for the remaining seven cells (in that part of the 16 cell table where $B_{N+1} = T$), then the rule is $0111111 = 127$. If a cellular automaton is operating under the model of Rule 127, the updated cell is colored white if the three relevant cells at the previous time step were all black. (A CA operating under Rule 127 is a series of alternating black and white horizontal lines.)

Can you guess what short formula captures Rule 127? If, as discussed above, all of the variable settings have a functional assignment of T except for the first one of ABC, then $f(A, B, C) = \overline{ABC}$. The unambiguous *Mathematica* syntax is **Not[And[p, q, r]]**.

This, as mentioned before for functions of two variables, is NOT the same as $f(A, B, C) = \overline{A}\,\overline{B}\,\overline{C}$. The unambiguous *Mathematica* syntax for this situation is **And[Not[p], Not[q], Not[r]]**.

For $f(A, B, C) = \overline{ABC} \equiv$ **Not[And[p, q, r]]**, we see that if at least one of A, B, or C is F, then the functional assignment is T. This function can be thought of as an analog to the NAND logic function. The $\uparrow$ binary operator also had a 0 in cell 1 of its joint probability table.

So there are eight functions with one 0 rotated through cells 1 to 8. There are $\binom{8}{2} = 28$ functions $f(A, B, C)$ with two 0s somewhere in the eight cells. There are $\binom{8}{3} = 56$ functions $f(A, B, C)$ with three 0s somewhere in the eight cells. Rule 110 was one of these 56 functions from the total of 256 where the three 0s are located in cells 1, 4, and 8. The greatest number of ways occurs, $\binom{8}{4} = 70$, when four 0s are distributed amongst the eight cells. There are thus 70 functions where the DNF expansion consists of four terms where the variable settings in those four terms take on the functional assignment of T.

What is an example? Arbitrarily, pick the four cells 1, 3, 5, and 7 of the joint probability table and insert a 0. This is the same as picking the four cells 2, 4, 6, and 8 and inserting a legitimate probability. Stick to our standard way of constructing the joint probability table and numbering the cells within the table. Thus, $B_{N+1} = T$ are the first eight cells, and $B_{N+1} = F$ are the last eight cells. This has the consequence that the following four joint probabilities will receive the non–zero assignments, say q:

1. $P(B_{N+1}A_N B_N \overline{C}_N) = q$

2. $P(B_{N+1}A_N \overline{B}_N \overline{C}_N) = q$

3. $P(B_{N+1}\overline{A}_N B_N \overline{C}_N) = q$

4. $P(B_{N+1}\overline{A}_N \overline{B}_N \overline{C}_N) = q$

Likewise, the following four joint probabilities will receive the zero assignments:

1. $P(B_{N+1}A_N B_N C_N) = 0$

2. $P(B_{N+1}A_N \overline{B}_N C_N) = 0$

3. $P(B_{N+1}\overline{A}_N B_N C_N) = 0$

4. $P(B_{N+1}\overline{A}_N \overline{B}_N C_N) = 0$

The eight cells under the $\overline{B}_{N+1}$ part of the joint probability table will contain the corresponding 0s and qs such that Bayes's Theorem will produce a black or white cell with certainty.

Construct a table like that shown below in Table 8.1. The expansion of 01010101 to base 10 determines that this is Rule 85. Using this table as a guide, match up

Table 8.1: *The functional assignment for three variables which is Rule 85.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
F	T	F	T	F	T	F	T
0	1	0	1	0	1	0	1

each T with 1 and a black cell, and match up each F with 0 and a white cell. Thus, a *wbw* pattern for three cells at time step N is the same as FTF leading to a T and a black cell at time step $N + 1$. The cell indexing $B_{N+1}\overline{A}_N B_N \overline{C}_N$ must have a q assigned with a 0 assigned to cell $\overline{B}_{N+1}\overline{A}_N B_N \overline{C}_N$. The four 0s are placed into cells 1, 3, 5, and 7 of the joint probability table.

Because of the symmetry of the binomial formula, there are 56 functions with five 0s, 28 functions with six 0s, and 8 functions with seven 0s. The final two functions, $\binom{8}{0} = 1$ and $\binom{8}{8} = 1$ are the two constant functions where either T or F is assigned to every variable possibility.

As the final double–check, add up all these functions with all possible zeroes placed somewhere within the joint probability table,

$$\sum_{k=0}^{8} \binom{8}{k} = 1 + 8 + 28 + 56 + 70 + 56 + 28 + 8 + 1 = 256 = 2^8$$

Exercise 8.6.11. Demonstrate the equivalency of two models with two eight cell joint probability tables with one model and a sixteen cell table. Base the models on Rule 110.

Solution to Exercise 8.6.11

For this example, suppose that there are four statements labeled generically as A, B, C, and D. The state of knowledge about any joint statement would be represented by one 16 cell joint probability table if the statements can only assume the values of TRUE or FALSE.

However, we emphasized in this Chapter that this situation could just as easily be thought about in terms of three statements A, B, and C under two models, D and $\overline{D}$, given by the notation $\mathcal{M}_1$ and $\mathcal{M}_2$. Thus, we can think in terms of two 8 cell joint probability tables with the first table's assignment under model $\mathcal{M}_1$, $P(ABC \mid \mathcal{M}_1)$, and the second table's assignments under model $\mathcal{M}_2$, $P(ABC \mid \mathcal{M}_2)$.

See Figure 8.3 below for the assignments under the two models. The two models were inspired by Rule 110 and its dual, so there are three zeroes in the top joint probability table where the DNF expansion for Rule 110 has a coefficient of F. The other five cells where the DNF expansion of Rule 110 has a coefficient of T have an assignment of 1/5.

Figure 8.3: *Two eight cell joint probability tables reflecting the assignments under two models.*

The second eight cell joint probability table in the bottom half of Figure 8.3 is the function that is dual to Rule 110. It has zeroes where its counterpart has an assignment of 1/5, and an assignment of 1/3 where it counterpart has zeroes. This pattern in the placement of the numerical assignments under each model is designed, of course, to enforce the deterministic behavior dictated by Rule 110 when implemented from the probabilistic standpoint.

Notice especially that the assignments sum to 1 under each model, or, in other words, the assignments in cells 1 through 8 sum to 1 given each model. Any one of the probabilities for a joint statement under the two models can be read directly

from the appropriate cell of each joint probability table. Thus,

$$P(A, B, C \mid \mathcal{M}_1) = 0 \text{ and } P(A, B, C \mid \mathcal{M}_2) = 1/3$$

Calculating $P(A, B, C)$ from the Bayesian modeling averaging perspective,

$$\sum_{k=1}^{\mathcal{M}} P(A, B, C, \mathcal{M}_k) = P(A, B, C)$$

$$
\begin{aligned}
P(A, B, C) &= \sum_{k=1}^{\mathcal{M}} P(A, B, C \mid \mathcal{M}_k) \, P(\mathcal{M}_k) \\
&= P(A, B, C \mid \mathcal{M}_1) \, P(\mathcal{M}_1) + P(A, B, C \mid \mathcal{M}_2) \, P(\mathcal{M}_2) \\
&= (0 \times 5/8) + (1/3 \times 3/8) \\
&= 1/8
\end{aligned}
$$

But how did we know the $P(\mathcal{M}_k)$? As we shall discover in the next Chapter, the answer just calculated is exactly the same as the one calculated under a one model, 16 cell joint probability table. That larger table must incorporate joint probabilities that include the two models. The assignments, except of course for the zero assignments, are different than for the two tables above. You may jump ahead and view such a joint probability table based on Rule 110 in Figure 9.1.

Nonetheless, the sum of the assignments over the sixteen cells must still equal 1. Each cell of the larger 16 cell joint probability table indexes a joint statement concerning not three statements, but four statements A, B, C, and D. Equivalently, the probability for the joint statement concerning A, B, and C, after marginalizing over D, is,

$$
\begin{aligned}
P(A, B, C) &= P(A, B, C, D) + P(A, B, C, \overline{D}) \\
&= 0 + 1/8 \\
&= 1/8
\end{aligned}
$$

If it hasn't become obvious in working through these exercises, I need to point out that careful attention must be paid to where the statement labels are placed in the construction of the joint probability table. Cell numbers index different joint statements depending on where statements A, B, C, and so on, together with their negations, appear. Compare Figures 8.1 and 8.3, for example.

The supplemental exercises delve into more detail about how the *Mathematica* function **Tuples[ ]** dictates the construction of a joint probability table.

Chapter 9

Probabilistic Cellular Automata

9.1   Introduction

With the events of the last few Chapters, we are now ready to address the issue of probabilistic cellular automata. These automata, as the Chapter title implies, are probabilistic generalizations of the deterministic one–dimensional cellular automata first examined in Chapter Three, and just recently discussed in the last Chapter.

We have already seen how probability generalizes Classical Logic. My personal canonical example is Jaynes's generalization of the classical *modus ponens* argument, $P(A \mid B, A \rightarrow B)$. A is an undecidable proposition if we use deduction. On the other hand, we do know something about A if we make an inference.

Before we proceed to similar generalizations involving CA, we have already demonstrated, not too surprisingly, that within this class of yet-to-be-discussed probabilistic cellular automata, some CA are *deterministic*. In fact, the class of 256 elementary CA described by Wolfram are deterministic. Simply put, the color of the cell to be updated is known with certainty.

However, if all we could demonstrate was that Bayes's Theorem reproduces the same CA output as a direct application of one of the 256 rules, the introduction of probability would be entirely superfluous. Worse still, it would be seen as a Rube Goldberg device cloaking matters in a complexity that added nothing of value.

The purpose of this Chapter is to advance the cause of making the transition from deduction to inference. Probability is the central linchpin for bridging this gap. Wolfram's CA provide excellent abstract examples for how information processing is involved in understanding the distinction between deduction and inference, or, perhaps better put, the distinction between an ontological model of reality and an information processor's knowledge about that reality.

It might be worthwhile to tarry a while longer with the deterministic CA before transitioning to probabilistic CA. Thus, we begin this Chapter with an extensive discussion of that most famous of Wolfram's deterministic CA, Rule 110. Rule 110 can apparently compute anything that *can be* computed. However, the rather embarrassing question of just what Rule 110 is actually computing the answer to must be glossed over since no one, including Wolfram, seems to know.

9.2   The Rule 110 Deterministic CA

We revisit the deterministic cellular automaton that follows Rule 110. But now we embed our understanding of that CA within the context of the manipulation rules for probability symbols, and especially within the context of Bayes's Theorem.

9.2.1   Bayes's Theorem for Rule 110

Using the notation introduced at the end of the last Chapter, let $\mathcal{M}_{110}$ stand for the model that implements Rule 110. Then, given this model guiding the evolution of the cellular automaton, together with the colors of three cells at the previous N^{th} time step, we can ask the question: What is the probability for the color of the cell that is ready to be updated at the next time step?

Start off by writing down the generic notation for a probability involving four statements, (not counting $\mathcal{M}_k$ since it will always remain to the right of the conditioned upon symbol as given),

$$P(D = T \,|\, A = T, B = T, C = T, \mathcal{M}_k)$$

In the notation for CA, this is expressed as,

$$P(B_{N+1} = b \,|\, A_N = b, B_N = b, C_N = b, \mathcal{M}_{110})$$

for the probability that the color of the cell to be updated is black, given that the cell above, B_N, and its two neighbors, A_N and C_N, were also black.

In addition to the colors of the three cells at the current time step, model $\mathcal{M}_{110}$ is placed to the right of the conditioned upon symbol as given. This indicates, once again, that Rule 110, and not some other logic function with three arguments, is assumed to be the actual rule governing the evolution of the CA. From the discussion in the last Chapter, Rule 110 serves as the inspiration for a single model $\mathcal{M}_{110}$ which places zero and non–zero numerical assignments into the sixteen cells of a joint probability table.

Filling in the template for Bayes's Theorem, we find that the required probability that the cell to be updated is black given that three relevant cells on the previous step were all black is expressed as,

$$P(B_{N+1} \,|\, A_N, B_N, C_N, \mathcal{M}_{110}) = \frac{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{110})}{P(A_N, B_N, C_N \,|\, \mathcal{M}_{110})}$$

This is the simplest version of Bayes's Theorem, and shows that the conditional probability on the left hand side is a ratio of two joint probabilities on the right hand side. The denominator, as we have seen, is a marginal probability that sums over all the possibilities for the cell that is to be updated. There are only two possibilities for any cell of the CA that is about to be updated; it can be either black, B_{N+1}, or white, $\overline{B}_{N+1}$.

Rewriting the denominator to include this fact, we now have,

$$P(B_{N+1} \,|\, A_N, B_N, C_N, \mathcal{M}_{110}) =$$

$$\frac{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{110})}{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{110}) + P(\overline{B}_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{110})}$$

The notation in Bayes's Theorem for other cell color situations is easy to understand. Remember that a black cell stands for T and a white cell for F. A line appearing above a statement variable indicates that the cell is colored white.

For example, the probability that the color of the updated cell is white given the cell above was white, the left neighbor was black and the right neighbor was white, is,

$$P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_{110}) =$$

$$\frac{P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_{110})}{P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_{110}) + P(B_{N+1}, A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_{110})}$$

As complicated as this looks, it is still just a conditional probability on the left hand side expressed on the right hand side as a ratio of a joint probability over a marginal probability, itself the sum of two joint probabilities.

9.2.2   A joint probability table for Rule 110

Just as in the last Chapter, it is easier to see what is going on by constructing a joint probability table. Figure 9.1 at the top of the next page shows just such a table consisting of four appropriately labeled variables and sixteen cells.

It is broken down into four tables of four cells each. The two tables on top are for the variable B_{N+1} where the updated cell is colored black, and the two tables on the bottom are for $\overline{B}_{N+1}$ where the updated cell is colored white. In addition to containing the numerical values of the probabilities, each cell is labeled from 1 to 16.

The numbers in the cells are seen to be legitimate numerical assignments to joint probabilities as might be made under some model. In fact, they are the ones that can be made under a model reflecting Rule 110. All of the many marginal probabilities are shown as well.

Figure 9.1: *Joint probability table for the Rule 110 elementary cellular automaton. The numbers placed in the cells make this a deterministic CA. All of the various marginal probabilities are shown as well.*

9.2.3   How certain is the color of a cell?

Here is a numerical example using the joint probability table. We want to check whether the probabilistic formulation will return the correct color of a cell in the cellular automaton with certainty, that is, with a probability of 1. Consider the example posed at the end of section 9.2.1.

$$P(w \mid b, w, w, \mathcal{M}_{110}) =$$

$$\frac{P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N \mid \mathcal{M}_{110})}{P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N \mid \mathcal{M}_{110}) + P(B_{N+1}, A_N, \overline{B}_N, \overline{C}_N \mid \mathcal{M}_{110})}$$

The notation on left hand side has been changed to make things a little bit more transparent. It clearly indicates that we are looking for the probability that the updated cell is white given that it was white at the previous time step and its two

neighbors were black and white. The right hand side retains the usual notation for variables set to T or F.

Find the cells in the joint probability table that correspond to the two joint probabilities on the right hand side of Bayes's Theorem. These are cells 12 and 4. Now substitute their numerical values to find,

$$P(w \,|\, b, w, w, \mathcal{M}_{110}) = \frac{1/8}{1/8 + 0} = 1$$

We calculate an answer that we had hoped for. It is certain that the updated cell will be colored white under Rule 110.

9.2.4 The DNF helps to fill in the joint probability table

Why are there numerical assignments of zero in a joint probability table? The reason is that the model we are using to drive the evolution of the CA does not allow that combination reflected by the joint statement where the 0s appear.

For example, there is a 0 in cell 1 of the joint probability table for Rule 110. Cell 1 is the joint statement that the cell to be updated is black and all three cells at the previous time step were also black. But Rule 110 does not permit this possibility. If all three cells at the previous time step were black, then the updated cell must be white. Refer back to the picture of Rule 110 in Figure 3.2 on pg. 68.

Back in Chapters Two and Three, we derived the full DNF for Rule 110. We will write down the DNF once again, this time rearranging the five terms to match the order of non–zero assignments in the cells of the joint probability table,

$$\mathcal{M}_{110} \equiv AB\overline{C} \vee A\overline{B}C \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$$

The DNF and Bayes's Theorem can then be used together as a heuristic device to determine which cells of the joint probability table should contain either zero or non–zero assignments.

In the last Chapter, we investigated a useful heuristic whereby the DNF for a three variable Boolean function could be used to help figure out the numerical assignments in a joint probability table for a four variable Boolean function. We employ that heuristic here to fill in the sixteen cells of the joint probability table for the deterministic Rule 110 cellular automaton.

Append B_{N+1} to the relevant cells at the previous time steps so that expressions like $B_{N+1}A_NB_NC_N$ and $B_{N+1}A_N\overline{B}_NC_N$ indicate two locations in the first eight cells of the sixteen cell joint probability table. These two cells contain the probability for a joint statement where the updated color is black. The terms in the Rule 110 DNF expansion given above tell us where the zero and non–zero numerical assignments must be made.

Likewise, cells in the bottom half of the joint probability table are indicated by expressions like $\overline{B}_{N+1}\overline{A}_NB_NC_N$ and $\overline{B}_{N+1}A_NB_N\overline{C}_N$ where the updated color is

white. Because Bayes's Theorem must return a 1 or a 0 for a deterministic system, the placement of the zero and non–zero numerical assignments in the second set of eight cells is also known.

For example, consider a particular case of Bayes's Theorem where the numerical value in cell 2,

$$B_{N+1} A_N B_N \overline{C}_N$$

will be the value in the numerator as well as the first term of the denominator. Then, we know that cell 10,

$$\overline{B}_{N+1} A_N B_N \overline{C}_N$$

will be the second term in the denominator. Therefore, if the conditional probability on the left hand side of Bayes' Theorem is to work out to a value of 1, this second term in the denominator must be 0. We can fill in cell 10 of the joint probability table with a 0, and cell 2 with some legitimate non–zero assignment. The $AB\overline{C}$ term in the DNF expansion told us that cell 2 must contain a non–zero assignment because the functional assignment for $f(T, T, F)$ was T.

The merit of using the DNF is now apparent. Each one of the five terms in the DNF, together with Bayes's Theorem, can be used to fill in all the cells of the joint probability table with legitimate numerical values that satisfy the constraints of the model.

The marginal values of the joint probability table are not without interest.[1] For example, the marginal probability for an updated cell to be black, irrespective of what has happened before, is $P(B_{N+1}) = 5/8$.

9.2.5   Summarizing through another example

Let's take stock of where we've been by looking at another example of a deterministic cellular automaton. We will gather together the same arguments and apply them in a more concise manner. But everything will be completely familiar from the Rule 110 template.

Consider another one of Wolfram's simple one–dimensional cellular automata. We select this CA, just like the Rule 110 CA, from his set of 256 possibilities. To repeat, these CA consist of cells that can be colored black or white, and which evolve by following a rule involving the cell above and its two neighbors at the previous time step. The CA chosen for this example follows Rule 30.

We bring over Bayes's Theorem almost intact. The only thing that needs to be changed is the model making the numerical assignments. Instead of conditioning on Rule 110, we now condition on the assumed truth of Rule 30. Thus, wherever $\mathcal{M}_{110}$ appeared before, we replace it with $\mathcal{M}_{30}$.

---

[1]In fact, these marginal probabilities assume a prominent role when the Maximum Entropy algorithm is used to assign numerical values according to some model.

This change in models has no impact whatsoever from the formal manipulation standpoint. The only (albeit important) implication of conditioning on the truth of Rule 30 is that the numerical assignments to joint probabilities will change.

Once again, use the color notation on the left hand side and the variable notation on the right hand side. The probability for some designated cell to be colored black given that it was black at the previous time step and its two neighboring cells were also colored black, is written via Bayes's Theorem as,

$$P(b \,|\, b, b, b, \mathcal{M}_{30}) = \frac{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{30})}{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{30}) + P(\overline{B}_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_{30})}$$

We decode Rule 30 to find the equivalent binary number. The resulting pattern of 1s and 0s in this binary number establishes the DNF representation for the Boolean function of three variables. With the DNF at our disposal, it becomes an easy matter to fill in the cells of a joint probability table for Rule 30.

Rule 30 as a binary number is,

$$\begin{aligned} 30 \ &= \ (0 \times 2^7) + (0 \times 2^6) + (0 \times 2^5) + (1 \times 2^4) + \\ &\quad (1 \times 2^3) + (1 \times 2^2) + (1 \times 2^1) + (0 \times 2^0) \\ &= \ 00011110 \end{aligned}$$

The pattern of 0s and 1s, when matched to the pattern of the ordered triples from the domain of the function, yields Table 9.1. The number of Ts shown in this

Table 9.1: *The functional assignment for three variables which is Rule 30.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
F	F	F	T	T	T	T	F

table as functional assignments will tell us how many terms will be in the DNF expansion. Thus, there will be four terms in the DNF. From the table, we readily reconstruct the DNF for Rule 30 as,

$$A\overline{B}\,\overline{C} \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$$

These four terms in the DNF, together with Bayes's Theorem, tell us where to insert the zero and non–zero assignments into the cells of the joint probability table. Figure 9.2 at the top of the next page shows such a joint probability table with the marginal probabilities filled in as well.

Now, finally, with these legitimate numerical assignments in the joint probability table, we can calculate the probability for the color of a cell that is scheduled to be updated via Rule 30.

B_{N+1}

A_N

	C_N	$\overline{C}_N$	
B_N	0 (1)	0 (2)	0
$\overline{B}_N$	0 (3)	1/8 (4)	1/8
	0	1/8	1/8

$\overline{A}_N$

	C_N	$\overline{C}_N$		
B_N	1/8 (5)	1/8 (6)	1/4	1/4
$\overline{B}_N$	1/8 (7)	0 (8)	1/8	1/4
	1/4	1/8	3/8	1/2

$\overline{B}_{N+1}$

A_N

	C_N	$\overline{C}_N$	
B_N	1/8 (9)	1/8 (10)	1/4
$\overline{B}_N$	1/8 (11)	0 (12)	1/8
	1/4	1/8	3/8

$\overline{A}_N$

	C_N	$\overline{C}_N$			
B_N	0 (13)	0 (14)	0	1/4	1/2
$\overline{B}_N$	0 (15)	1/8 (16)	1/8	1/4	1/2
	0	1/8	1/8	1/2	

1/4	1/4	1/2	1/4	1/4	1/2
		1/2	1/2		1.00

Figure 9.2: *Joint probability table for the Rule 30 cellular automaton. The numbers placed in the cells make this a deterministic CA.*

$$
\begin{aligned}
P(b \mid b, b, b, \mathcal{M}_{30}) &= \frac{P(B_{N+1}, A_N, B_N, C_N \mid \mathcal{M}_{30})}{P(B_{N+1}, A_N, B_N, C_N \mid \mathcal{M}_{30}) + P(\overline{B}_{N+1}, A_N, B_N, C_N \mid \mathcal{M}_{30})} \\[2mm]
&= \frac{0}{0 + 1/8} \\[2mm]
&= 0
\end{aligned}
$$

It is certain that the cell will not be colored black; it will be colored white. Thus, this purely probabilistic result corresponds to applying Rule 30 in the usual deterministic manner. The numerical values appearing in Bayes's Theorem for this case are contained in cells 1 and 9 of the joint probability table for Rule 30.

It is noteworthy that the marginal probabilities for all four variables considered separately, B_{N+1}, A_N, B_N, and C_N, are all 1/2 under this model. Interestingly, Wolfram implemented a random number generator using this CA. Rule 30 happens to be an example of a model where all the interesting action takes place through

the intermediation of the joint effects of the variables. These are simply called "interactions" and they play a vital role in developing models that capture in some sense the notion of "complexity," "surprise," "accident," or "unpredictable."[2]

9.3 Generalizing Deterministic Cellular Automata

Just as probability theory generalized Classical Logic, so it will perform the same valuable service for deterministic CA. We will (temporarily) label the CA produced by this kind of generalization of the deterministic cellular automata as *probabilistic cellular automata*. Later, we will discuss why such a term is an oxymoron.

9.3.1 Losing information about the previous time step

Previously, we saw that probabilistic inferences could be made even when some of the information necessary for a deduction went missing. For example, we quibbled at calling variable A undecidable when the `OR` logical operator was the given model and the setting for variable B was the only information provided.

We adjusted our mental framework to accommodate a state of knowledge where A could be thought of being logically implicated as T or F, not with certainty, but rather with a numerical assignment of probability equal to, say, $1/2$.

Consider another example much discussed, the logic operator `CONDITIONAL`. The setting for A is required in determining the truth of B when the classical implication logic operator is used. However, in another generalization afforded by probability, this *modus ponens* type logical deduction was not thought to be invalid when only B was provided and assumed TRUE. Our state of knowledge that A must be TRUE was not undecidable, but rather was determined to either stay the same or increase.

What if the color of only one cell at the previous time step is known, instead of the color of all three cells? This situation is analogous to the above examples in that a previously deterministic CA would now have to be treated probabilistically to arrive at an answer.

Suppose that we return to a CA governed by Rule 110. The right neighbor of the cell to be updated was colored black at the previous time step. That is all we know. What can we surmise about the color of the cell at the next time step?

Before, when the color of the cell at the previous time step and its two neighbors were provided, we knew with certainty what the color would be. Now we cannot be certain, but this does not imply that all is lost. We can still apply Bayes's Theorem in the same way to update our state of knowledge given what *has* been provided.

---

[2]In case it escaped your attention, I purposefully chose not to include the word "random." Remember that Rule 30 reflects a deterministic system.

Letting a $\star$ stand in for the now missing information about both the color of the cell to be updated at the previous time step as well as its left neighbor, Bayes's Theorem is written as,

$$P(b \,|\, \star, \star, b, \mathcal{M}_{110}) = \frac{P(B_{N+1}, C_N \,|\, \mathcal{M}_{110})}{P(C_N \,|\, \mathcal{M}_{110})}$$

The numerator and denominator in Bayes's Theorem are now marginal probabilities.

Referring back to the joint probability table for Rule 110, sum over the relevant cells to find the marginal probabilities required. The numerator, $P(B_{N+1}, C_N)$, is the sum over the four cells, cells 1, 3, 5, and 7, while the denominator, $P(C_N)$, is sum over eight cells, cells 1, 3, 5, 7, 9, 11, 13, and 15.

Thus,

$$
\begin{aligned}
P(b \,|\, \star, \star, b, \mathcal{M}_{110}) &= \frac{0 + 1/8 + 1/8 + 1/8}{0 + 1/8 + 1/8 + 1/8 + 1/8 + 0 + 0 + 0} \\[2mm]
&= 3/4
\end{aligned}
$$

Now, as remarked, we can no longer be certain that the cell is black given only that its right neighbor was black at the previous time step. But what we are entitled to know, despite the dearth of information, is that the truth of the statement that the cell is black is supported by the probability measure to the tune of 3/4. If one looks at all eight updating possibilities provided under Rule 110, one finds that in the four cases where the right neighbor is black, three of these result in the updated cell also being colored black.

9.3.2   Losing information about the model in control

There is another way in which deterministic CA can be generalized by probability. Above, we saw what would happen when we lost information about the color of some cells at the previous time step. However, the rule governing the evolution of the CA was never in question. But what would happen if even this was subject to some uncertainty? What if we lost information about which rule was driving the evolution of the CA?

We have derived some formal manipulation rules in Chapter Seven to cover this situation. These rules confirm what our unaided intuition would speculate is the state of knowledge if there were some uncertainty about the rule running the CA.

In the simplest case, imagine that we know that *two* rules are the only candidates behind the observed CA. Suppose that these two rules are our familiar Rule 110 and Rule 30, but we happen to be completely ignorant as to which one of these two it might be.

What is the IP's state of knowledge that the updated cell is white given that the cell was white at the previous times step and its two neighbors were black and

white? One formula we could use to arrive at an answer is,

$$P(\overline{B}_{N+1} \mid A_N, \overline{B}_N, \overline{C}_N) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}_{N+1} \mid A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_k) \, P(\mathcal{M}_k \mid A_N, \overline{B}_N, \overline{C}_N)$$

The answer is in the form of an average. It is the average of the result from Bayes's Theorem under each model. The averaging is done with respect to the state of knowledge about the two rules in question when these models are conditioned on what is known. In this case, what is known are the colors of all three relevant cells at the previous time step.

From previous discussions of Rule 110, we know that it updates a cell to the color white given the status at the previous time step. So,

$$P(w \mid b, w, w, \mathcal{M}_{110}) = 1$$

Rule 30, on the other hand, updates a cell to the color black under these conditions. So,

$$P(w \mid b, w, w, \mathcal{M}_{30}) = 0$$

Presumably, our state of knowledge about either model should not be affected by knowing the colors of the cells at the previous time step. After all, these conditions reflect one legitimate setting of the three arguments to either logic function. So this cannot serve to distinguish between them. If the IP started out in a state of total ignorance about which rule was operative, it must remain in this same state of total ignorance. Thus,

$$P(\mathcal{M}_{110} \mid b, w, w) = P(\mathcal{M}_{30} \mid b, w, w) = 1/2$$

In this beginning example, we postulated only $\mathcal{M} = 2$ models, labeled as $\mathcal{M}_{110}$ and $\mathcal{M}_{30}$. Both are deterministic models inspired by Rule 110 and by Rule 30. And, if we are completely at a loss about which of these two rules is actually driving the CA, we assign a probability of $1/2$ to each model. Thus, the probability for the updated cell to be colored white, given that it was colored white at the previous time step and its two neighbors were black and white, using the above averaging formula, is,

$$P(w \mid b, w, w) = (1 \times 1/2) + (0 \times 1/2) = 1/2$$

You have every right to be suspicious of verbal arguments like this one. Confirm the tedious details in Exercises 9.5.15 and 9.5.16.

The IP resides in a state of knowledge that is maximally uncertain about whether the updated cell is colored black or white even though it knows the disposition of the determining cells at the previous time step. Apparently, something needs to be done to update the IP's state of knowledge about the two models in question in order to make any better predictions.

Thus, there exist certain equivalencies in our state of knowledge that arise in different ways. We just learned that lack of knowledge about which of two models are controlling the CA resulted in maximal uncertainty about the color of the cell to be updated even when we knew what the causal factors were.

But the same situation would arise if we knew for certain that Rule 30 was the only rule operative, but didn't know the color of any of the cells at the previous time step. Our state of knowledge is the same, $P(w \mid \star, \star, \star, \mathcal{M}_{30}) = 1/2$. To verify this, refer back to the joint probability table for Rule 30 in Figure 9.2. The marginal probability for $P(\overline{B}_{N+1} \mid \mathcal{M}_{30}) = 1/2$.

9.3.3   Losing information about the deterministic model

There is yet another way in which we are able to generalize the deterministic CA via probability. Casual observation of the joint probability tables for both Rule 110 and Rule 30 reveals that 0s will always appear somewhere. These 0s are legitimate numerical assignments to joint probabilities and are, of course, mandated by the deterministic nature of the elementary CA in question. The 0s allowed Bayes's Theorem to calculate a probability for the color of an updated cell to be 1 or 0.

There will be appropriately placed 0s in some cells for every deterministic model. Recall that in the case of the logic functions, 0s also had to be placed into the appropriate cells of their respective joint probability tables. This would seem to follow from the Boolean Algebra perspective because deterministic CA were simply the extrapolation from the domain of two variables to the domain of three variables with the carrier set remaining fixed at T and F.

Figure 9.3 exhibits a different scenario. We have now relaxed this just discussed requirement about 0s by placing legitimate numerical assignments other than 0 in *all* of the cells of the joint probability table for a generalized CA. Now, the probability for the color of the updated cell will not be 1 or 0, even when conditioned on just one given model.

This could be argued as a very mild departure from the joint probability tables for the deterministic CA. Now some model is assigning a legitimate numerical value of 1/16 to each of the 16 cells of the table. Figure 9.3 illustrates in detail the construction of such a table. Label such an assignment as model $\mathcal{M}_{\star}$.

The probability for any cell to be updated to black given that the three relevant cells at the previous time step were also all black now becomes

$$P(b \mid b, b, b, \mathcal{M}_{\star}) = \frac{1/16}{1/16 + 1/16} = 1/2$$

We are taking advantage of the right to express Bayes's Theorem in this case simply as the ratio of the value in cell 1 over the addition of the values in cells 1 and 9.

B_{N+1}

	A_N				$\overline{A}_N$				
	C_N	$\overline{C}_N$			C_N	$\overline{C}_N$			
B_N	1/16 (1)	1/16 (2)	1/8		1/16 (5)	1/16 (6)	1/8	1/4	
$\overline{B}_N$	1/16 (3)	1/16 (4)	1/8		1/16 (7)	1/16 (8)	1/8	1/4	
	1/8	1/8	1/4		1/8	1/8	1/4	1/2	

$\overline{B}_{N+1}$

	A_N				$\overline{A}_N$				
	C_N	$\overline{C}_N$			C_N	$\overline{C}_N$			
B_N	1/16 (9)	1/16 (10)	1/8		1/16 (13)	1/16 (14)	1/8	1/4	1/2
$\overline{B}_N$	1/16 (11)	1/16 (12)	1/8		1/16 (15)	1/16 (16)	1/8	1/4	1/2
	1/8	1/8	1/4		1/8	1/8	1/4	1/2	
	1/4	1/4	1/2		1/4	1/4	1/2		
					1/2	1/2			1.00

Figure 9.3: *Joint probability table for a cellular automaton. The numbers placed in the cells make this a probabilistic CA, not a deterministic CA. Please take note that the descriptive adjectives "probabilistic" and "deterministic" for a CA are an oxymoron and redundant, respectively.*

Obviously, we can no longer be certain that the updated cell is black or white as we could when one of the deterministic CA was the given model. The state of knowledge captures the IP's ignorance about the color of the updated cell given the model $\mathcal{M}_\star$ reflected in the new joint probability table.

Notice that the marginal probabilities for this assignment are as uninformative as possible. The marginal probabilities over eight cells like $P(B_{N+1})$ are all equal to 1/2, the marginal probabilities over four cells like $P(\overline{B}_{N+1}, C_N)$ are all equal to 1/4, and the marginal probabilities over two cells like $P(\overline{A}_N, B_N, C_N)$ are all equal to 1/8.

9.4   Connections to the Literature

It is very important for me to point out that probabilistic CA as I have described them here are NOT the way Wolfram described them (see pp. 155–160, pp. 591–592, and pg. 922 in [20]). The resolution of this issue is extremely critical because it involves subtle *conceptual* errors concerning probability.

First of all, Wolfram mentions *continuous* CA whose cells can have any color on a continuous gray scale instead of just black and white. He talks about these continuous CA in the same breath as probabilistic CA. He tells us that they are basically the same thing. Allowing the gray scale to be resolved on a finer and finer level is the same as allowing the color of the cell to be represented by some number between 0 and 1.

In other words, each cell can be assigned as black or white "with some fixed independent probability p." He then shows visual examples of the same starting CA evolving in different ways following this precept of assigning color probabilistically.

But from the general perspective of Boolean Algebra, a carrier set $\mathbf{B}$ may consist of more than just the two elements T and F. And a functional assignment from $\mathbf{B}$ can then be something other than T or F. There is no probability or inferencing involved in any of this. It is all strictly part of the deductive, deterministic system that is Boolean Algebra.

We presented a simple example of this back in section 3.6.2 and Exercise 3.7.11. The elementary CA were generalized by allowing a cell to be colored light gray or dark gray in addition to black and white. The evolution of this CA could be calculated from some Boolean formula involving two new elements in the carrier set. For example, let a and a' correspond to light gray and dark gray, just as T corresponded to black and F to white. There was no need to drag probability into the fray even though we had increased the resolution of the gray scale.

Everything within this generalized CA proceeded according to its status as an ontological system. An information processor knew everything about this system. Therefore, there was no need to perform an inference as opposed to a deduction.

What is the hallmark of the correct conceptual grasp of what probability is all about? It is in understanding that probability refers not to the ontological status of an object, but rather to an information processor's epistemological knowledge about that object. This fundamental conceptual distinction is the *sine qua non* marking the division between clarity and confusion.

We are considering deterministic CA as convenient ontological models of the real world. For us, this is merely a very interesting, abstract, general, easy, and again, a convenient starting point so that further arguments may proceed. As an aside, Wolfram takes his CA very seriously indeed as ontological models of the real world, having made preliminary arguments that the very structure of space and time, as well as the ultimate Final Theory of Physics, should best be thought of as some kind of advanced CA, perhaps some sort of causal network.

We took great pains in this Chapter to emphasize that it is the information reflected in an IP's model (or models) that allows it to make inferences whenever deductions could not be made. The IP might be missing information about, say, the color of relevant cells at previous time steps, or the particular rule that is governing the evolution of the CA. Just as in the generalization of Classical Logic, even if the IP *is* missing information, it can still make inferences based on the information that it *does* possess.

Because the IP lacks information in the models that *it* is using does not mean that the ontology of the real world is in any way affected by such a defect. Suppose the world really is running according to Rule 110. Probability does NOT therefore enter into some reformulation of Rule 110. Probability does not enter in any way whatsoever at the fundamental level of the ontological operations.

We must disagree at a fundamental conceptual level with Wolfram's description that the color of the cell *is determined by some probability*. The color of the cell is determined by the rule. The IP may not know with certainty the color of that cell because it is missing crucial information in the models it is employing. The IP therefore uses probability and inferencing to attach a degree of belief in the cell's color. This is the only thing that it can do under these circumstances.

Therefore, one cannot say that any CA rules can be probabilistically altered, or that the color of the cell is *probabilistically determined*. The juxtaposition of these two words is an oxymoron.

The rules are what they are, and they are completely understood as ontological entities. They are deterministic. There is no missing information. It is only *our* lack of information that causes us to introduce probabilities. Rule 110 can not be probabilistically altered into a probabilistic CA at the ontological level.

Of course, to say all of this is to utter one of the great shibboleths of our scientific age. Quantum mechanics has dictated that probability must be present in the very ontological foundations of reality. I can only speculate that Wolfram's thorough imbibing of quantum orthodoxy spilled over into his explanation of these so–called probabilistic CA.

But like Laplace, Einstein, Schrödinger, Jaynes, and many others, my mind recoils at just what it could possibly mean to say that something at the ontological level is *probabilistically determined*. On the other hand, I have no problem accepting that the finite capabilities of any IP may be inadequate for a full understanding of complex ontological systems. As a result, its *knowledge* about these systems will be forever couched in probabilistic terms. In the end, we all have to accept Kant's dictum that the *Ding an sich* is ultimately unknowable.

9.5   Solved Exercises for Chapter Nine

Exercise 9.5.1. Why is there a 0 in cell 8 of the joint probability table for Rule 110 as shown in Figure 9.1?

Solution to Exercise 9.5.1

Cell 8 indexes the joint statement, $B_{N+1}\overline{A}_N\,\overline{B}_N\,\overline{C}_N$. In words, this statement could be expressed as, "The updated cell is black and the colors of the three relevant cells at the previous time step were all white."

However, the fundamental Boolean formula that encapsulates Rule 110 states that the functional assignment of F is given to the variable setting of $A = F$, $B = F$, and $C = F$. It simply can not happen, under this function, that T is the functional assignment.

Translated into the language of cellular automata, when the CA is following Rule 110 it is impossible for the updated cell to be colored black given that its three predecessor cells were all colored white. Refer back to Figure 3.2 to confirm this. Therefore, a 0 must be inserted into the joint probability table that indexes this joint statement. The sequence of events described by this joint statement cannot take place under a model following Rule 110.

Exercise 9.5.2. How many rules like Rule 30 are there?

Solution to Exercise 9.5.2

With this question, we are asking more precisely how many rules like Rule 30 possess four 0s somewhere in the first eight cells of the joint probability table. There are, in fact, $\binom{8}{4} = 70$ rules for elementary cellular automata where one might find four 0s as assigned numerical values to probabilities of joint statements.

Exercise 9.5.3. Use the decomposition into a binary number to find the functional assignment of Rule 150. Construct a table like Table 9.1.

Solution to Exercise 9.5.3

The decimal number 150 is decomposed into the binary number 10010110. Rule 150 is seen to be one of the 70 rules like Rule 30 where four 0s are assigned as numerical values to the probabilities of joint statements. Table 9.2 illustrates the functional assignment to three variables based on this particular binary number.

Table 9.2: *The functional assignment for three variables which is Rule 150.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
T	F	F	T	F	T	T	F

Exercise 9.5.4. What is the DNF representation for Rule 150?

Solution to Exercise 9.5.4

Based on an examination of Table 9.2, the DNF for Rule 150 would consist of the four terms corresponding to the functional assignment of T.

$$ABC \lor A\overline{B}\,\overline{C} \lor \overline{A}B\overline{C} \lor \overline{A}\,\overline{B}C$$

Exercise 9.5.5. Wolfram has a different looking expression for Rule 150. Is our DNF representation as given above and Wolfram's representation logically equivalent?

Solution to Exercise 9.5.5

Wolfram, ([20], page 864), gives this expression for Rule 150 when translated into the notation of Chapter Two,

$$\text{Rule } 150 \ \equiv (A \oplus B) \oplus C$$

Logical equivalency takes place when the `EQUAL` function returns T for the two expressions in question. The logic expression,

$$((A \oplus B) \oplus C) \leftrightarrow (A \land B \land C) \lor (A \land \overline{B} \land \overline{C}) \lor (\overline{A} \land B \land \overline{C}) \lor (\overline{A} \land \overline{B} \land C)$$

must return T for all eight possible combinations of the variables.

We happily accept the trade–off of writing a small *Mathematica* program (refer to the Appendices) versus working out Boolean operations by hand. Construct the function that implements the proposed logical equivalency as,

```
logicExpression[p_, q_, r_] := Equivalent[Xor[Xor[p, q], r],
                Or[And[p, q, r], And[p, Not[q], Not[r]],
        And[Not[p], q, Not[r]], And[Not[p], Not[q], r]]]
```

Mathematica's tautology testing function,

```
TautologyQ[logicExpression[p, q, r]]
```

returns **True** for all eight possible variable settings, thus confirming that the two ways of writing Rule 150 are one and the same.

Exercise 9.5.6. In what cell numbers of the 16 cell joint probability table would the 0s for Rule 150 be placed?

Solution to Exercise 9.5.6

Four 0s in cell numbers 2, 3, 5, and 8, and four 0s in cell numbers 9, 12, 14, and 15.

Exercise 9.5.7. Does such a placement of 0s reproduce the correct color for an updated cell of a CA following Rule 150?

Solution to Exercise 9.5.7

Suppose we are curious as to whether the updated cell of the CA following Rule 150 is white given that the previous relevant cells were colored black, white, and black. Cast this question into a form so that it can be answered by Bayes's Theorem,

$$P(w \mid b, w, b, \mathcal{M}_{150}) = \frac{P(\overline{B}_{N+1} A_N \overline{B}_N C_N)}{P(\overline{B}_{N+1} A_N \overline{B}_N C_N) + P(B_{N+1} A_N \overline{B}_N C_N)}$$

Substitute the correct cell numbers into the right hand side,

$$P(w \mid b, w, b, \mathcal{M}_{150}) = \frac{\text{Cell } 11}{\text{Cell } 11 + \text{Cell } 3}$$

From the last exercise, we learned that cell 3 must be 0 and cell 11 a non–zero assignment. Therefore,

$$P(w \mid b, w, b, \mathcal{M}_{150}) = \frac{q}{q + 0} = 1$$

Thus, for a deterministic CA it is certain that the updated cell will be colored white under the specified conditions. Or, saying the same thing, it is certain the updated cell will not be colored black, $P(b \mid b, w, b, \mathcal{M}_{150}) = 0$.

Confirm this result by looking at Table 9.2 in order to verify that the functional assignment of F is made to the variable settings of $A = T$, $B = F$, and $C = T$.

Exercise 9.5.8. Rule 150 is the model governing the evolution of a CA. Suppose that the information processor has lost information about both the color of a cell and its right neighbor at the previous time step. What is the information processor's state of knowledge about black being the color of the cell scheduled for updating?

Solution to Exercise 9.5.8

Say that the known color of the cell to the left of the cell to be updated was white at the previous time step. Then, the information processor's state of knowledge

about the color of the cell to be updated is,

$$P(b \mid w, \star, \star, \mathcal{M}_{150}) = \frac{P(B_{N+1}\overline{A}_N \mid \mathcal{M}_{150})}{P(\overline{A}_N \mid \mathcal{M}_{150})}$$

Exercise 9.5.9. What cell numbers from the joint probability table for Rule 150 are involved in calculating the answer to Bayes's Theorem as given in the previous exercise?

Solution to Exercise 9.5.9

The probability for the joint statement in the numerator $B_{N+1}\overline{A}_N$ is a marginal probability over the four cells, 5, 6, 7, and 8. The probability for the joint statement in the denominator $\overline{A}_N$ is a marginal probability over the eight cells, 5, 6, 7, 8, 13, 14, 15, and 16.

Exercise 9.5.10. What is the information processor's state of knowledge about the updated cell having the color black given the described loss of information?

Solution to Exercise 9.5.10

From Exercise 9.5.6, we know where the 0s are in the joint probability table. Assume that the non–zero assignments have been given the same value q. Then, in the numerator, cells 5 and 8 are 0, so cells 6 and 7 add up to $2q$. In the denominator, cells 5, 8, 14, and 15 are 0, so cells 6, 7, 13, and 16 add up to $4q$. Thus,

$$P(b \mid w, \star, \star, \mathcal{M}_{150}) = \frac{2q}{4q} = 1/2$$

The information processor is maximally uncertain about the color of the updated cell when restricted to such meager information. Note that in the last four columns of Table 9.2 where the variable A has value F, two of the four functional assignments are T.

Exercise 9.5.11. What if the information processor thought that either Rule 150 or Rule 110 was governing the evolution of the CA, but was uncertain which one it was?

Solution to Exercise 9.5.11

We are still under the supposition that the information processor is burdened with the previously described lack of information about two of the relevant cells. To complicate things, there is additional uncertainty about which of two models is appropriate. Use the formula presented in section 9.3.2 to calculate the state of knowledge.

For the first term in the summation we need the answer from Bayes's Theorem under two different models, $P(b \,|\, w, \star, \star, \mathcal{M}_{110})$ and $P(b \,|\, w, \star, \star, \mathcal{M}_{150})$. We've just finished finding the answer for Rule 150, so we'll concisely show the answer for Rule 110.

In this case, the numerical assignments come from the joint probability table shown in Figure 9.1. The same cells are involved as discussed above in the previous exercises.

$$
\begin{aligned}
P(b \,|\, w, \star, \star, \mathcal{M}_{110}) &= \frac{P(B_{N+1}\overline{A}_N \,|\, \mathcal{M}_{110})}{P(\overline{A}_N \,|\, \mathcal{M}_{110})} \\[2mm]
&= \frac{3/8}{4/8} \\[2mm]
&= 3/4
\end{aligned}
$$

We are still maximally uncertain about the two models even knowing that a white cell was the left neighbor at the previous time step. Thus,

$$
\begin{aligned}
P(b \,|\, w, \star, \star) &= P(b \,|\, w, \star, \star, \mathcal{M}_{110}) \, P(\mathcal{M}_{110} \,|\, \overline{A}_N) + P(b \,|\, w, \star, \star, \mathcal{M}_{150}) \, P(\mathcal{M}_{150} \,|\, \overline{A}_N) \\[2mm]
&= (3/4 \times 1/2) + (1/2 \times 1/2) \\[2mm]
&= 5/8
\end{aligned}
$$

The information processor is slightly more certain that the updated cell is black when these two models are considered. It is somewhat counter–intuitive that, being in some sense more uncertain at the start with two rules in play, there is less uncertainty about the color when compared to the uncertainty when only the one model, Rule 150, was being considered. But the probability for a black update was higher under model $\mathcal{M}_{110}$, so the average prediction from these two models results in increased certainty about the outcome.

Exercise 9.5.12. Show in more detail the marginal probabilities for the joint probability table shown as Figure 9.3 as discussed at the end of section 9.3.3.

Solution to Exercise 9.5.12

The marginal probabilities are sums over the relevant cells in the full 16 cell joint probability table. The three marginal probabilities mentioned were,

$$
P(B_{N+1}) = \sum_{A_N, B_N, C_N}^{8} P(B_{N+1}, A_N, B_N, C_N)
$$

$$P(\overline{B}_{N+1}, C_N) \;=\; \sum_{A_N, B_N}^{4} P(\overline{B}_{N+1}, A_N, B_N, C_N)$$

$$P(\overline{A}_N, B_N, C_N) \;=\; \sum_{B_{N+1}}^{2} P(B_{N+1}, \overline{A}_N, B_N, C_N)$$

The notation at the bottom of the summation symbol indicates that the variables in question are to be summed over all of their possible values. If a variable is not referenced, then it remains fixed at its value as the argument.

Since each variable takes on only two values, black or white, A_N, B_N, C_N at the bottom of the summation symbol for $P(B_{N+1})$ dictates a summation over $2^3 = 8$ cells of the joint probability table. B_{N+1} remains fixed as black in the summation.

Thus, the first marginal probability is the sum over cells 1 through 8, the second is the sum over cells 9, 11, 13, and 15, while the third is the sum over cells 5 and 13. For this last marginal probability, the sum is explicitly the sum of two terms,

$$P(\overline{A}_N, B_N, C_N) = P(B_{N+1}, \overline{A}_N, B_N, C_N) + P(\overline{B}_{N+1}, \overline{A}_N, B_N, C_N)$$

Exercise 9.5.13. What is the marginal probability of $P(B_N, \overline{C}_N)$?

Solution to Exercise 9.5.13

Drawing on the results of the last exercise, we can write,

$$P(B_N, \overline{C}_N) = \sum_{B_{N+1}, A_N}^{4} P(B_{N+1}, A_N, B_N, \overline{C}_N)$$

We are going to sum over the four possibilities for B_{N+1} and A_N while B_N and $\overline{C}_N$ remain fixed at black and white.

$$P(B_N, \overline{C}_N) = P(B_{N+1}, A_N, B_N, \overline{C}_N) + P(B_{N+1}, \overline{A}_N, B_N, \overline{C}_N)$$

$$+ P(\overline{B}_{N+1}, A_N, B_N, \overline{C}_N) + P(\overline{B}_{N+1}, \overline{A}_N, B_N, \overline{C}_N)$$

This is the sum over cells 2, 6, 10, and 14.

Exercise 9.5.14. Discuss how model $\mathcal{M}_{110}$ can be conceptualized as a four variable Boolean function.

Solution to Exercise 9.5.14

The carrier set is $\mathbf{B} = \{T, F\}$ and functions of four variables are denoted by,

$$f : \mathbf{B}^4 \to \mathbf{B}$$

An arbitrary Boolean function with four arguments might be written either as $f(x_1, x_2, x_3, x_4)$, or as $f(w, x, y, z)$.

Expand the arbitrary Boolean function with orthonormal basis functions and coefficients as,

$$f(w, x, y, z) =$$

$$[f(T, T, T, T) \circ w \circ x \circ y \circ z] \bullet [f(T, T, T, F) \circ w \circ x \circ y \circ z'] \bullet \cdots$$

$$\bullet [f(F, F, F, F) \circ w' \circ x' \circ y' \circ z']$$

with the expansion on the right hand side consisting of sixteen terms in all.

Since the carrier set consists of only T and F, any function assignment to terms like $f(T, T, T, F)$ must also be T or F. Thus, the function expansion will collapse into a collection of terms looking either like $T \circ w \circ x \circ y \circ z'$ or $F \circ w \circ x \circ y \circ z'$. In the first case, the term is reduced to $w \circ x \circ y \circ z'$, or is eliminated in the second case.

For the particular situation involving $\mathcal{M}_{110}$, examine each term in the expansion one by one. Translate the notation of the first term, $w \circ x \circ y \circ z$, into $B_{N+1} A_N B_N C_N$. The updated cell B_{N+1} cannot be black under Rule 110 if the three relevant cells at the previous time step were also all black. Thus, this term is eliminated because $f(T, T, T, T) = F$.

Look at the second term, $w \circ x \circ y \circ z'$ and translate to $B_{N+1} A_N B_N \overline{C}_N$. Can the updated cell be black under Rule 110 if the three relevant cells were black, black and white? Yes, it can. Thus, $f(T, T, T, F) = T$ and the second term $w \circ x \circ y \circ z'$ is retained in the overall expansion of the function $f(w, x, y, z)$.

Proceeding in this manner, the first eight terms in the expansion reduce to the five terms where $f(\star, \star, \star, \star) = T$,

$$f(w, x, y, z) = wxyz' \bullet wxy'z \bullet wx'yz \bullet wx'yz' \bullet wx'y'z \bullet \cdots?$$

Because of the symmetry involved, it is easy to figure out the second eight terms. Wherever $f(\star, \star, \star, \star) = T$ for any of the first eight terms, it now equals $f(\star, \star, \star, \star) = F$ for the second eight terms, and vice versa. Thus, the ninth term, $w' \circ x \circ y \circ z$ corresponds to $\overline{B}_{N+1} A_N B_N C_N$. Can the updated cell be white if the three previous cells were all black? Yes, it must be white under Rule 110. Thus, this term will be included in the expansion because $f(F, T, T, T) = T$.

In all, there will be three more terms where $f(\star, \star, \star, \star) = T$ in the second set of eight terms. They are,

$$w'xyz \bullet w'xy'z' \bullet w'x'y'z'$$

The full expansion is thus,

$$f(w, x, y, z) = wxyz' \bullet wxy'z \bullet wx'yz \bullet wx'yz' \bullet wx'y'z \bullet w'xyz \bullet w'xy'z' \bullet w'x'y'z'$$

This expression can be reduced through standard Boolean operations to an even shorter expression consisting of just five terms,

$$f(w, x, y, z) = wx'y \bullet wyz' \bullet wy'z \bullet w'y'z' \bullet w'xyz$$

So, as a test of this answer, what is the functional assignment to $f(F, T, F, T)$? Plug in the values for the four arguments $w, x, y,$ and z. Apply the Boolean operators $\circ$ and $\bullet$ on the right hand side showing that the functional assignment for this four variable Boolean function mimicking $\mathcal{M}_{110}$ is $f(F, T, F, T) = F$. What is the probability of an updated white cell given a black, white, and black cell at the previous time step? That is, what is $P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N)$? By Rule 110, a black, white, and black cell produce a black cell. Therefore, $P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N) = 0$ confirming the formula for the Boolean function.

Table 9.3 summarizes the four variable Boolean function just found by showing the functional assignment for all sixteen possible settings for the four arguments. The first two rows correspond to the way we described Rule 110 by a three variable Boolean function in Chapter Three. The last two rows present the symmetrical arrangement demanded by deterministic CA and earlier called the dual function. We see the five Ts in the second row and the three Ts in the last row that played a role in our explanation of the function expansion.

Table 9.3: *The functional assignment table for the four variable Boolean function mimicking Rule 110.*

$TTTT$	$TTTF$	$TTFT$	$TTFF$	$TFTT$	$TFTF$	$TFFT$	$TFFF$
F	T	T	F	T	T	T	F
$FTTT$	$FTTF$	$FTFT$	$FTFF$	$FFTT$	$FFTF$	$FFFT$	$FFFF$
T	F	F	T	F	F	F	T

In Wolfram's numbering scheme, this Boolean function is the 28,305th among all possible four variable Boolean functions because the binary number,

$$0110\,1110\,1001\,0001$$

matching the functional assignment of Ts with 1s and Fs with 0s expands to the decimal number 28,305.

A small *Mathematica* expression,

```
BooleanTable[BooleanFunction[28305, 4]]
```

provides confirmation since it outputs the list,

```
{False, True, True, False, True, True, True, False,
   True, False, False, True, False, False, False, True}
```

matching Table 9.3.

Exercise 9.5.15. Verify the answer in section 9.3.2 from first principles.

Solution to Exercise 9.5.15

Writing out Bayes's Theorem in the simplest version we have,

$$P(\overline{B}_{N+1} \mid A_N, \overline{B}_N, \overline{C}_N) = \frac{P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N)}{P(A_N, \overline{B}_N, \overline{C}_N)}$$

As usual, the denominator is a sum over the two values for B_{N+1},

$$P(A_N, \overline{B}_N, \overline{C}_N) = \sum_{B_{N+1}}^{2} P(B_{N+1}, A_N, \overline{B}_N, \overline{C}_N)$$

But we haven't accounted for the presence of the two models.

$$P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N) = \sum_{k=1}^{2} P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_k)$$

Thus, the numerator will consist of two terms and the denominator four terms,

$$P(\overline{B}_{N+1} \mid A_N, \overline{B}_N, \overline{C}_N) = \frac{\sum_{k=1}^{2} P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_k)}{\sum_{k=1}^{2} \sum_{B_{N+1}}^{2} P(B_{N+1}, A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_k)}$$

The number of statements inside the probability expressions on the right hand side has increased from four to five. Previously, we conditioned on a single model, and thus required only four statements. Conceptually, we are dealing with one 32 cell joint probability table in this way of writing out Bayes's Theorem.

The numerical assignments over all 32 cells of this one joint probability table must sum to 1. Therefore, the numerical assignments become 1/16 wherever they were 1/8 in Figures 9.1 and 9.2. The two values in the numerator come from cell 12 from model $\mathcal{M}_{110}$ and cell 28 from model $\mathcal{M}_{30}$. The four values in the denominator come from cells 12, 28, 4, and 20.

$$P(w \mid b, w, w) = \frac{1/16 + 0}{1/16 + 0 + 1/16 + 0} = 1/2$$

confirming our quick answer in section 9.3.2.

Exercise 9.5.16. **Verify the answer in section 9.3.2 from the formula actually given there.**

Solution to Exercise 9.5.16

The formula given in section 9.3.2 is a version of Bayes's Theorem that is commonly used in data analysis. The predictions from each model are averaged with respect to the probability of the model after taking account of any observed data.

Here, we don't have any data *per se*, but the causal factors serve the same role for updating the relative status of the models under consideration. Repeating the formula,

$$P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, \overline{C}_N) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, A_N, \overline{B}_N, \overline{C}_N)$$

We have already determined that model $\mathcal{M}_{110}$ outputs an updated white cell, while model $\mathcal{M}_{30}$ outputs an updated black cell given that the causal factors are the relevant three cells at the previous time step. The first term has therefore been determined as 1 and 0 under each model.

The second term is,

$$P(\mathcal{M}_k \,|\, A_N, \overline{B}_N, \overline{C}_N) = \frac{P(A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{k=1}^{2} P(A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}$$

where $P(A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_k)$ is expanded into,

$$P(A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_k) = P(B_{N+1}, A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_k) + P(\overline{B}_{N+1}, A_N, \overline{B}_N, \overline{C}_N \,|\, \mathcal{M}_k)$$

The prior probability for each model, $P(\mathcal{M}_k)$, when the IP is in a state of total ignorance is,

$$P(\mathcal{M}_{110}) = P(\mathcal{M}_{30}) = 1/2$$

Substituting the numerical assignments under each model, we have,

$$P(\mathcal{M}_1 \,|\, A_N, \overline{B}_N, \overline{C}_N) = \frac{(0 + 1/8) \times 1/2}{[\,(0 + 1/8) \times 1/2\,] + [\,(1/8 + 0) \times 1/2\,]} = 1/2$$

Thus, $P(\mathcal{M}_2 \,|\, A_N, \overline{B}_N, \overline{C}_N) = 1/2$ as well, and,

$$P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, \overline{C}_N) = (1 \times 1/2) + (0 \times 1/2) = 1/2$$

verifying the answer found by the other approach used in the last exercise.

Chapter 10

Logic Puzzles

10.1 Introduction

This is a short Chapter that serves more as a brief interlude. Or, if you are in a charitable mood, you might take it as possessing some entertainment value. We want to stop for a moment to see if we can solve some typical "logical puzzles," or "brain–twisters."

For many people, myself included, their first exposure to any kind of logical reasoning was in the form of these entertaining puzzles. Somewhat surprisingly, they turned out to be more difficult than anticipated, and impervious to quick solution.

After some preliminary head scratching, invariably followed by doodling and the odd sketch, any attempt at resolving the puzzle was quickly abandoned. It is no wonder that most people came to the conclusion that as wonderful as "logical reasoning" might be, it was something their poor brains were incapable of handling.

But with the systematic development of information processing as begun in this Volume, the realization of the "trick" behind many of these puzzles finally emerges. We will reveal the mechanical steps involved in deciphering these "brain–twisters." This discovery will fit quite nicely into the scheme of things as developed so far.

We look at two puzzles of this type as presented by Brown [3]. He shows how to solve such puzzles through what he calls "syllogistic reasoning." Interesting as this deductive approach is, we prefer to apply probabilistic reasoning as a generalization of logical reasoning.

To that end, we go ahead and translate Brown's solutions into our standard framework for inferential reasoning and, not unexpectedly, the joint probability table plays a prominent role in such a framework.

10.2   Alfred Goes to College

Here is the first and easier of the two puzzles. Alfred is a student at college with the following set of three constraints on his life. If Alfred studies, then he receives good grades. If Alfred doesn't study, then he enjoys college. If Alfred doesn't receive good grades, then he doesn't enjoy college. What may we logically conclude about Alfred's life?

The punch line, after you have scratched your head for a while, is that Alfred definitely receives good grades. Really? Is one entitled to come to such a conclusion? There is definitely some lingering doubt.

The probabilistic framework starts with statements in a state space. Each such statement is either TRUE or FALSE. Here are the three relevant statements in the Alfred puzzle together with the notation for the corresponding logical variable.

1. $E \equiv$ "Alfred enjoys college."

2. $G \equiv$ "Alfred gets good grades."

3. $S \equiv$ "Alfred studies."

The obvious interpretation of the overbar for each variable becomes,

1. $\overline{E} \equiv$ "Alfred does not enjoy college."

2. $\overline{G} \equiv$ "Alfred does not get good grades."

3. $\overline{S} \equiv$ "Alfred does not study."

A joint probability table for these three statements each taking on just two possible values consists of $2^3 = 8$ cells. The joint probability table is sketched out in Figure 10.1 showing the three statements, and with the eight cells properly numbered. Each cell indexes a probability that the joint statement is TRUE under some model.

For example, cell 5 is the container for the probability that the joint statement, "Alfred does not enjoy college and Alfred gets good grades and Alfred studies." is TRUE under some model. In symbols, this is written as $P(\overline{E}SG \,|\, \mathcal{M}_k)$.

The model $\mathcal{M}_k$ assigning legitimate numerical values to all eight cells of the joint probability table is derived from the constraints listed in the puzzle. All three of these constraints are in the form of an implication, $A \rightarrow B$. Collected together, these implications are,

1. $S \rightarrow G$

2. $\overline{S} \rightarrow E$

3. $\overline{G} \rightarrow \overline{E}$

Figure 10.1: *An eight cell joint probability table for solving the Alfred logic puzzle.*

In words, the constraints of the puzzle are that,

1. If Alfred studies, then he gets good grades.

2. If Alfred doesn't study, then he enjoys college.

3. If Alfred doesn't get good grades, then he doesn't enjoy college.

From our previous work, we know that the logic function of two arguments called `IMPLIES` is written generically as,

$$f_{13}(A, B) \equiv A \rightarrow B$$

and can be expanded into the full DNF consisting of three terms,

$$f_{13}(A, B) = AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$$

In turn, this means that the function $\phi_2(A, B) = A\overline{B}$ from the full set of orthonormal functions in the expansion template,

$$f_j(A, B) = \sum_{k=1}^{4} c_k(A, B)\ \phi_k(A, B)$$

must have the coefficient $c_2(A, B) = F$. Therefore, in a joint probability table, as we have seen before, the cell for $P(A\overline{B})$ has a numerical assignment of 0.

So now we are making some progress because the three implications listed above will allow us to place 0s into appropriate cells of the joint probability table. For

example, the very first constraint in the model is that $S \rightarrow G$, or $P(S\overline{G}) = 0$. But expanding $S\overline{G}$ according to the standard Boolean operations yields

$$P(S\overline{G}) \equiv P(ES\overline{G} \vee \overline{E}S\overline{G}) = 0$$

Thus, a 0 will be placed into cells 3 and 7 of the joint probability table.

Do exactly the same thing for the remaining two constraints specified in the puzzle.

$$\overline{S} \quad \rightarrow \quad E$$

$$P(\overline{S}\,\overline{E}) \quad = \quad 0$$

$$P(\overline{E}\,\overline{S}G \vee \overline{E}\,\overline{S}\,\overline{G}) \quad = \quad 0$$

0s will be placed into cells 6 and 8 from this constraint.

$$\overline{G} \quad \rightarrow \quad \overline{E}$$

$$P(\overline{G}E) \quad = \quad 0$$

$$P(ES\overline{G} \vee E\overline{S}\,\overline{G}) \quad = \quad 0$$

0s will be placed into cells 3 and 4 from this constraint, but a 0 had already been placed into cell 3 by the first constraint.

All together there are five 0s appearing in cells 3, 4, 6, 7, and 8. Three of the eight cells, cells 1, 2, and 5, will have some legitimate non–zero numerical assignment. Arbitrarily choose a numerical assignment of $1/3$ for each of these three cells because the exact value doesn't make any difference as far as the whole point of the puzzle is concerned.

Figure 10.2 at the top of the next page shows the joint probability table filled in with all of these numerical assignments under the model dictated by the constraints in the Alfred puzzle. Every possible inference or logical deduction that we care to make is now immediately calculable from Bayes's Theorem.

But we don't even have to go that far to confirm the punch line of the puzzle. It is immediately evident that $P(\overline{G}) = 0$ or that $P(G) = 1$ by summing over the relevant cells in the joint probability table to find the marginal sum. It is certain that Alfred receives good grades.

Did Alfred enjoy college given that it was a fact that he studied and he received good grades? We will find out in just a second that, colloquially speaking, we don't know if he enjoyed college under these conditions. To answer any kind of general query of this nature, set up Bayes's Theorem,

Figure 10.2: *The joint probability table for the Alfred logic puzzle filled in with numerical assignments that satisfy some model. The model implements the information in the statement of the puzzle.*

$$P(E \mid S, G, \mathcal{M}_k) = \frac{P(ESG \mid \mathcal{M}_k)}{P(ESG \mid \mathcal{M}_k) + P(\overline{E}SG \mid \mathcal{M}_k)}$$

$$= \frac{1/3}{1/3 + 1/3}$$

$$= 1/2$$

Contrary to the certain knowledge that Alfred received good grades, the IP's state of knowledge about whether Alfred enjoyed college reflects missing information in the model. The IP's degree of belief in whether Alfred enjoyed college hovers halfway between believing it to be TRUE and believing it to be FALSE, even though he studied and got good grades.

Now it is very, very important, continuing our discussion at the end of the last Chapter, to emphasize that Alfred himself does not exist in some superposition of states where he halfway enjoyed college and halfway did not enjoy college.[1] If we ask him, he will give us a definite response that he did or did not enjoy college. It is simply that *we*, the information processors, don't know whether he did or not enjoy college. The model making our numerical assignments did not provide us with enough *information* to make that call.

---

[1] We didn't want Stephen Hawking to go looking for his gun, so we refrained from mentioning *Schrödinger's Cat*.

10.3   The Halloween Party

Now we discuss a second, and harder, "brain–twister" involving who will attend a Halloween party. There are four protagonists in this puzzle, Alice, Ben, Charlie, and Diane. Because of some apparent personality conflicts among our little group, there are some constraints governing who will show up at the party. These constraints are:

1. If Alice goes, then Ben won't go and Charlie will.

2. If Ben and Diane go, then either Alice or Charlie (but not both) will go.

3. If Charlie goes and Ben does not, then Diane will go, but Alice will not.

Who will go to the party and under what conditions?

   If there is any merit to my claim that we can apply a mechanical procedure to these types of puzzles, then we are going to have to set up another joint probability table. And, the three constraints in the puzzle are going to dictate the placement of some zeroes in this joint probability table so that some definite conclusions can be made.

   If I were to proceed in a linear fashion and dissect the problem in detail, as done for the first puzzle, I would rapidly lose your attention. Therefore, let's relegate most of the fascinating details to the exercises, and get to the solution as quickly as possible.

   There are four statements in the state space indicating whether any of our four protagonists attended the party. There are only two options for each person, attend or don't attend, so there will be $2^4 = 16$ cells in the joint probability table. Figure 10.3 shows the full joint probability table with the numerical assignments already filled in under some model $\mathcal{M}_k$.

   The ten 0s are there because of the implications in the statement of the puzzle. As in the previous puzzle, we simply split up the probability evenly among the remaining six cells that were not ruled out by the information in the model.

   What is an immediately obvious conclusion to the puzzle from inspection of the joint probability table? Find the marginal probability for A by summing across cells 1 through 8, leading to $P(A \,|\, \mathcal{M}_k) = 0$. It is certain that Alice will not attend the party. Could you have figured that out on your own?

   Will Charlie go to the party if Ben and Diane go? To answer this query, set up Bayes's Theorem as,

$$P(C \,|\, B, D, \mathcal{M}_k) = \frac{P(BCD \,|\, \mathcal{M}_k)}{P(BD \,|\, \mathcal{M}_k)}$$

$$= \frac{\text{cell 1 } + \text{ cell 9}}{\text{cell 1 } + \text{ cell 9 } + \text{ cell 3 } + \text{ cell 11}}$$

Figure 10.3: *The joint probability table for the Halloween party logic puzzle filled in with numerical assignments that satisfy some model. The model implements the information in the statement of the puzzle.*

$$\text{cell } 1 = P(ABCD \,|\, \mathcal{M}_k)$$

$$\text{cell } 9 = P(\overline{A}BCD \,|\, \mathcal{M}_k)$$

$$\text{cell } 3 = P(AB\overline{C}D \,|\, \mathcal{M}_k)$$

$$\text{cell } 11 = P(\overline{A}B\overline{C}D \,|\, \mathcal{M}_k)$$

$$P(C \,|\, B, D, \mathcal{M}_k) = \frac{0 + 1/6}{0 + 1/6 + 0 + 0}$$

$$= 1$$

It is certain that Charlie will go to the party if Ben and Diane end up attending. Could you have figured that out on your own?

10.4   Solved Exercises for Chapter Ten

Exercise 10.4.1. Write out the symbolic logic expressions for each of the three verbally expressed constraints in the Halloween party puzzle.

Solution to Exercise 10.4.1

1. "If Alice goes, then Ben won't go and Charlie will."
   $A \rightarrow [\, \overline{B} \wedge C \,]$

2. "If Ben and Diane go, then either Alice or Charlie (but not both) will go."
   $[\, B \wedge D \,] \rightarrow [\, (\overline{A} \wedge C) \vee (A \wedge \overline{C}) \,]$ or $[\, B \wedge D \,] \rightarrow [\, A \oplus C \,]$

3. "If Charlie goes and Ben does not, then Diane will go but Alice will not."
   $[\, \overline{B} \wedge C \,] \rightarrow [\, \overline{A} \wedge D \,]$

Exercise 10.4.2. Write out equivalent logic expressions for each of the three constraints in the Halloween party puzzle to discover where the 0s will be placed among the sixteen cells of the joint probability table.

Solution to Exercise 10.4.2

The expressions we want are, in some sense, the opposite of the expansions where we keep the terms with a coefficient of T. Now we want to keep the coefficient of terms with F because we are trying to find out where the 0s are located. See Exercise 10.4.9 for more explanation.

$$A \rightarrow [\, \overline{B} \wedge C \,] \quad \equiv \quad AB \vee A\overline{C}$$

$$[\, B \wedge D \,] \rightarrow [\, (\overline{A} \wedge C) \vee (A \wedge \overline{C}) \,] \quad \equiv \quad ABCD \vee \overline{A}B\overline{C}D$$

$$[\, \overline{B} \wedge C \,] \rightarrow [\, \overline{A} \wedge D \,] \quad \equiv \quad A\overline{B}C \vee \overline{B}C\overline{D}$$

Exercise 10.4.3. Expand out the first and third expressions in the last exercise to include all four variables.

Solution to Exercise 10.4.3

We want to expand out to all four variables so we will know exactly which cells in the joint probability table must contain a 0. After expanding, the resulting expression may be simplified because some of the terms will be repetitions. The first constraint expands out to eight terms,

$$AB \vee A\overline{C} \quad \equiv \quad ABCD \vee AB\overline{C}D \vee ABC\overline{D} \vee AB\overline{C}\,\overline{D} \vee$$

$$AB\overline{C}D \vee A\overline{B}\,\overline{C}D \vee ABC\overline{D}\,\overline{D} \vee A\overline{B}\,\overline{C}\,\overline{D}$$

The third constraint expands out to four terms,

$$A\overline{B}C \vee \overline{B}C\overline{D} \equiv A\overline{B}CD \vee A\overline{B}C\overline{D} \vee \overline{A}\,\overline{B}C\overline{D} \vee \overline{A}\,\overline{B}C\overline{D}$$

The second constraint was already expressed in two terms of four variables.

Exercise 10.4.4. There are now 14 terms in four variables. Drop all the repeated terms in four variables.

Solution to Exercise 10.4.4

There are four repeated terms,

1. $AB\overline{C}D$

2. $AB\overline{C}\,\overline{D}$

3. $ABCD$

4. $A\overline{B}C\overline{D}$

leaving ten terms in four variables. It is easiest to find the repeated terms by matching up the terms with the appropriate cell number in the joint probability table. Thus,

1. The first expansion corresponds to cell numbers 1 2 3 4 3 7 4 8

2. The second expansion corresponds to cell numbers 1 and 11

3. The third expansion corresponds to cell numbers 5 6 6 14

illustrating that cells 1, 3, 4, and 6 are repeated. Here are the ten cells remaining, 1 2 3 4 5 6 7 8 11 14. These are the ten cells which will have a numerical assignment of 0 as shown in Figure 10.3.

Exercise 10.4.5. Examine some of the symbolic expansions together with their corresponding verbal constraints to see if they make sense.

Solution to Exercise 10.4.5

The full expansion of the first constraint, $A \rightarrow [\,\overline{B} \wedge C\,]$,

"If Alice goes, then Ben won't go and Charlie will."

has just been worked out to six non–duplicated terms involving four variables where each term must be FALSE. These six terms, each representing some joint statement, are,

$$ABCD \vee AB\overline{C}D \vee ABC\overline{D} \vee AB\overline{C}\,\overline{D} \vee A\overline{B}\,\overline{C}D \vee A\overline{B}\,\overline{C}\,\overline{D}$$

Look at the first term, $ABCD$. This is the joint statement, "Alice, Ben, Charlie, and Diane will all go to the party."

But this is FALSE under the first constraint because if Alice goes then Ben won't go. Look at the last term, $A\overline{B}\,\overline{C}\,\overline{D}$. This is the joint statement, "Alice goes to the party but Ben, Charlie and Diane don't go." Once again, under the first constraint if Alice goes, then Ben won't go which is OK, but then Charlie must go which is contradicted by the joint statement.

The full expansion of the second constraint, $[\,B \wedge D\,] \to [\,(\overline{A} \wedge C) \vee (A \wedge \overline{C})\,]$,

"If Ben and Diane go, then either Alice or Charlie (but not both) will go."

was shown to equal these two terms in four variables that must be FALSE,

$$ABCD \vee \overline{A}B\overline{C}D$$

The first term duplicates one from the first constraint. Look at the second term, $\overline{A}B\overline{C}D$. This is the joint statement, "Alice and Charlie don't go while Ben and Diane attend the party." Ben and Diane attended the party in this joint statement, but neither Alice nor Charlie were there, violating the second constraint that at least one of them should be there.

The full expansion of the third constraint, $[\,\overline{B} \wedge C\,] \to [\,\overline{A} \wedge D\,]$,

"If Charlie goes and Ben does not, then Diane will go but Alice will not."

had four terms, but one of them was a duplicate. So the three terms in four variables that must be FALSE because of this final restriction are,

$$A\overline{B}CD \vee A\overline{B}C\overline{D} \vee \overline{A}\,BC\overline{D}$$

Look at the first term, $A\overline{B}CD$. This is the joint statement, "Alice, Charlie and Diane go to the party, but Ben does not." So both Charlie and Diane attend the party, while Ben does not go and all of this is allowed. However, Alice also attends the party which is forbidden by the final constraint. So this joint statement is, in fact, FALSE as we hoped.

Exercise 10.4.6. Use Bayes's Theorem to verify Alice's non–attendance at the party under any circumstances.

Solution to Exercise 10.4.6

We already know that Alice will not attend the Halloween party under any set of conditions whatsoever. But just check to see what Bayes's Theorem reports about the state of knowledge concerning Alice's attendance, if, in fact, Ben was at the party while Charlie and Diane were not.

It is a given that Ben was at the party while Charlie and Diane were not. So place these statements to the right of the conditioned upon symbol. What we currently don't know is if Alice attended, so the IP wants to calculate in general,

$$P(A\,|\text{ any given conditions})$$

Here we will use Bayes's Theorem to find out about $P(A\,|B,\overline{C},\overline{D},\mathcal{M}_k)$ where $\mathcal{M}_k$ encapsulates all of the information in the three constraints, and therefore dictates what numerical assignments will be made to all the cells in the joint probability table.

$$
\begin{aligned}
P(A\,|B,\overline{C},\overline{D},\mathcal{M}_k) \;&=\; \frac{P(A,B,\overline{C},\overline{D}\,|\,\mathcal{M}_k)}{P(B,\overline{C},\overline{D}\,|\,\mathcal{M}_k)}\\[2mm]
&=\; \frac{\text{cell } 4}{\text{cell } 4 + \text{cell } 12}\\[2mm]
&=\; \frac{0}{0 + 1/6}\\[2mm]
&=\; 0
\end{aligned}
$$

As expected, the IP's state of knowledge indicates that it is impossible for Alice to have attended the party under these circumstances.

Of course, the first eight cells of the joint probability table all contain 0s. So the probability of any joint statement beginning with A will be 0, and Bayes's Theorem will report back that it was impossible for Alice to attend the party under any circumstances.

Exercise 10.4.7. What can you say about Ben's attendance if Charlie and Diane both attended the party?

Solution to Exercise 10.4.7

Here we see the value in having a mechanical (or algorithmic) method at our disposal that we can apply across the board in solving these logic puzzles. Strictly speaking, we can't say anything about Ben's attendance in this situation under a deductive framework. On the other hand, under an inferential approach, we can use Bayes's Theorem, together with the numerical assignments dictated by the model, to provide some quantitative answer to this question.

$$
P(B\,|\,C,D,\mathcal{M}_k) = \frac{P(B,C,D\,|\,\mathcal{M}_k)}{P(C,D\,|\,\mathcal{M}_k)}
$$

$$
= \frac{\text{cell } 1 + \text{cell } 9}{\text{cell } 1 + \text{cell } 9 + \text{cell } 5 + \text{cell } 13}
$$

$$= \frac{0 + 1/6}{0 + 1/6 + 0 + 1/6}$$

$$= 1/2$$

The IP is maximally uncertain about Ben's attendance. Its degree of belief in the statement, "Ben attended the party.", or the statement "Ben did not attend the party." is the same. Ben, in fact, did or did not attend the party. The information available to the IP in the form of the model that was employed was insufficient to disambiguate this fact.

The inferential approach at least provides a quantitative answer about the IP's degree of belief. It doesn't give up right at the start and claim that the problem is undecidable. It says rather, "Give me better information and I might be able to give you a better inference."

Exercise 10.4.8. Alice will not attend the party under any circumstances. Does that fact affect the previous result?

Solution to Exercise 10.4.8

No, the previous calculation implicitly took that fact into account. We can explicitly include it as a known fact and everything works out to the same answer.

$$
\begin{aligned}
P(B \,|\, \overline{A}, C, D, \mathcal{M}_k) &= \frac{P(\overline{A}, B, C, D \,|\, \mathcal{M}_k)}{P(\overline{A}, B, C, D \,|\, \mathcal{M}_k) + P(\overline{A}, \overline{B}, C, D \,|\, \mathcal{M}_k)} \\[2mm]
&= \frac{\text{cell } 9}{\text{cell } 9 + \text{cell } 13} \\[2mm]
&= \frac{1/6}{1/6 + 1/6} \\[2mm]
&= 1/2
\end{aligned}
$$

Exercise 10.4.9. Use the formal manipulation rules from Boolean Algebra to verify the transformations in Exercise 10.4.2.

Solution to Exercise 10.4.9

We already know that another way of expressing the implication logic function,

$$A \rightarrow B \text{ is equivalent to } \overline{A} \vee B$$

This, however, is where the functional assignment takes on the value T. We want the functional assignment F so that we can place 0s as the numerical assignments to

the proper cells of the joint probability table. Using **De Morgan's axiom**, change $\overline{A} \vee B = T$ to $A \wedge \overline{B} = F$ whence, of course, $P(A\overline{B}) = 0$.

Tentatively taking this result as a template, conjecture that the analog for the first constraint $A \rightarrow [\, \overline{B} \wedge C \,]$ is,

$$A \wedge (B \vee \overline{C}) \equiv AB \vee A\overline{C}$$

Expanding out these two terms as was done in Exercise 10.4.3 leads to the joint statements where the 0s must be placed.

Do the same thing for the second and third constraints. The third constraint is easier, so we work that one out first as,

$$[\, \overline{B} \wedge C \,] \rightarrow [\, \overline{A} \wedge D \,] \equiv (\overline{B} \wedge C) \wedge (A \vee \overline{D}) \equiv A\overline{B}C \vee \overline{B}C\overline{D}$$

The second constraint works out to,

$$
\begin{aligned}
[\, B \wedge D \,] \rightarrow [\, (\overline{A} \wedge C) \vee (A \wedge \overline{C}) \,] \;\; &\equiv \;\; B \wedge D \wedge [\, (A \vee \overline{C}) \wedge (\overline{A} \vee C) \,] \\[2mm]
&\equiv \;\; B \wedge D \wedge [\, A\overline{A} \vee AC \vee \overline{C}\,\overline{A} \vee \overline{C}C \,] \\[2mm]
&\equiv \;\; B \wedge D \wedge [\, AC \vee \overline{A}\,\overline{C} \,] \\[2mm]
&\equiv \;\; ABCD \vee \overline{A}B\overline{C}D
\end{aligned}
$$

Remember that on the right hand side of the equivalency symbol are the terms in the expansion with a coefficient of F, not a coefficient of T, because we are interested in finding where the zeroes should be located in the joint probability table.

Exercise 10.4.10. Can we verify consequences from the given premises?

Solution to Exercise 10.4.10

Brown [3, pg. 142] shows that his technique of "syllogistic reasoning" is able to verify consequents from the given premises of a logic puzzle. But we can do the same thing in an easier fashion by just referring back to the joint probability table. Brown asks us if the following statement is a consequence of the Halloween party puzzle.

> "If Alice and Ben both go to the party, or if neither of them goes, then Diane will go or Charlie will not go."

Symbolically, this consequence is,

$$[\, (A \wedge B) \vee (\overline{A} \wedge \overline{B}) \,] \rightarrow [\, D \vee \overline{C} \,]$$

But, as we have seen, by applying standard Boolean operations this is the same as,

$$C \wedge \overline{D} \wedge (\overline{A}\,\overline{B} \vee AB) \equiv \overline{A}\,\overline{B}C\overline{D} \vee ABC\overline{D}$$

Each of these two terms corresponds to some cell in the joint probability table. Both cells should contain 0s. Of course, the second term, $ABC\overline{D}$, starts with A and is therefore among the first eight cells. It must have a 0 as a numerical assignment.

The first term, $\overline{A}\,\overline{B}C\overline{D}$, corresponds to cell 14. This cell also has 0 as a numerical assignment under the model.

Thus, we have verified the above conjecture about our four Halloween party participants. Depending on how you want to look at it, it is either a consequence of the premises, or of the constraints. Or, in our preferred mode of inferential thinking, it is a consequence of the information contained in the model for the Halloween party puzzle.

Exercise 10.4.11. Is it possible to set up Bayes's Theorem to answer the question of whether Alfred enjoyed college conditioned on the fact that he did not get good grades?

Solution to Exercise 10.4.11

No, it is not. The reason why is the prohibition against placing a statement that is always FALSE to the right of the conditioned upon symbol. Any statement placed to the right of the conditioned upon symbol is assumed to be TRUE. Therefore, Bayes's Theorem will always have a problem with this situation, but will alert us to this fact by trying to divide by 0.

The conditional probability expression $P(E\,|\,\overline{G}, \mathcal{M}_k)$ is not valid because $\overline{G}$ is assumed TRUE, whereas under the information provided by the assumed model $\mathcal{M}_k$, it is always FALSE. Set up Bayes's Theorem anyway as,

$$P(E\,|\,\overline{G}, \mathcal{M}_k) = \frac{P(E\overline{G}\,|\,\mathcal{M}_k)}{P(\overline{G}\,|\,\mathcal{M}_k)} = \frac{0 + 0}{0} \qquad \text{(INVALID)}$$

to see the warning signal of a division by 0.

Chapter 11

Formal Rules for Prediction

11.1   Introduction

The goal here is to establish a state of knowledge about something that will happen in the future. By definition, such an event has not yet taken place. Presumably, as information processors, we are uncertain about how that event will turn out. Basically, what we are talking about is an information processor's ability to predict future events based on what is known about the past. Also needed are models to explain how both the future and the past depend on some set of causal factors.

For example, the color of the cell in one of Wolfram's 256 elementary CA at time step $N + 1$ will be determined after the relevant information processing involving the three cells at time step N takes place. Since these are deterministic CA, we know with certainty what color the updated cell will be.

On the other hand, as we hinted at in previous Chapters, if the information processor has to work with missing information, uncertainty will be introduced. Perhaps the particular model driving the evolution of the CA is unknown.

An information processor's state of knowledge is always defined operationally as a probability distribution. A probability distribution about some statement A_{N+1} one step into the future will be called a predictive probability distribution. The prediction depends on what has happened in the past. It depends as well on the whole gamut of models that have been entertained as possible explanations for the events.

It is easier to provide the first examples for these formal rules by predicting coin tosses and rolls of dice. The application to other events, and eventually to probabilistic cellular automata, will take place later.

11.2   The Prediction Formula

The derivation for the generally applicable formal rule begins with a joint probability over these three items:

1. the future event to be predicted,

2. the past events that have already taken place, and,

3. the models driving the assignment of legitimate numbers between 0 and 1 to the abstract symbols representing the joint probabilities.

   For ease of presentation, adopt the following terminology. A_2 is the statement at time step $N = 2$, that is, it is the statement about the future for which we would like to make a prediction. A_1 is the known outcome for the statement at time step $N = 1$, that is, the known outcome that has already taken place. And $\mathcal{M}_k$ is the kth model, that is, the particular model that inserts the actual numbers into a joint probability table. Assume that a total of $\mathcal{M}$ models will be considered.

   The joint probability in question then is $P(A_2, A_1, \mathcal{M}_k)$. The goal is to derive, through the auspices of the formal manipulation rules, the predictive probability distribution, $P(A_2 \,|\, A_1)$. What is the information processor's state of knowledge about the future event given what has happened in the past?

11.2.1   The derivation of the formula

To begin, use the **Sum Rule** to marginalize out the effects of all $\mathcal{M}$ models,

$$\sum_{k=1}^{\mathcal{M}} P(A_2, A_1, \mathcal{M}_k) = P(A_2, A_1)$$

Secondly, after invoking the **Commutativity axiom**, decompose the terms under the summation by the **Product Rule** into,

$$\sum_{k=1}^{\mathcal{M}} P(A_2, \mathcal{M}_k, A_1) = \sum_{k=1}^{\mathcal{M}} P(A_2 \,|\, \mathcal{M}_k, A_1) \, P(\mathcal{M}_k \,|\, A_1) \, P(A_1)$$

At this point we have,

$$P(A_2, A_1) = \sum_{k=1}^{\mathcal{M}} P(A_2 \,|\, \mathcal{M}_k, A_1) \, P(\mathcal{M}_k \,|\, A_1) \, P(A_1)$$

But $P(A_1)$ is a constant value that does not depend upon $\mathcal{M}_k$, so it can be brought outside the summation.

$$P(A_2, A_1) = P(A_1) \sum_{k=1}^{\mathcal{M}} P(A_2 \,|\, \mathcal{M}_k, A_1) \, P(\mathcal{M}_k \,|\, A_1)$$

The next step divides both sides by the constant value of $P(A_1)$ to yield,

$$\frac{P(A_2, A_1)}{P(A_1)} = \sum_{k=1}^{\mathcal{M}} P(A_2 \,|\, \mathcal{M}_k, A_1) \, P(\mathcal{M}_k \,|\, A_1)$$

But the left hand side is the definition of a conditional probability by Bayes's Theorem,

$$P(A_2 \,|\, A_1) = \frac{P(A_2, A_1)}{P(A_1)}$$

Thus we can write,

$$P(A_2 \,|\, A_1) = \sum_{k=1}^{\mathcal{M}} P(A_2 \,|\, \mathcal{M}_k, A_1) \, P(\mathcal{M}_k \,|\, A_1)$$

The final step in the derivation consists in the extremely important observation that it is the model, and only the model, that determines the numerical value attached to a statement. Any past data are completely irrelevant. Past data do not assign values to joint probabilities; only models are allowed to do that.

For example, if it were to happen that the model of a fair coin is the only model under consideration, then the probability for HEADS on the next toss is $1/2$ as dictated by that model. That probability is completely independent of how many HEADS and TAILS have appeared on any number of previous tosses.

Thus, we can eliminate A_1 from the first term on the right hand side to arrive at the predictive formula,

$$P(A_2 \,|\, A_1) = \sum_{k=1}^{\mathcal{M}} P(A_2 \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, A_1) \tag{11.1}$$

As alluded to before, the probability for the statement to be predicted is an average over the predictions made by each model. The averaging is done with respect to the probability allocated for each model as modified by its dependence on the past data A_1.

11.2.2   Predicting the first occurrence

It is interesting to see what this formula says about the very first occurrence of statement A. This is the state of affairs before any past trials have taken place, so there are no previous data to provide information about the relative standing of the proposed models.

Therefore, it is assumed that all of the models are on the same footing. That is, no one model is known to be better than any of its competitors due to some past data. Under this stipulation, the predictive formula in Equation (11.1) simplifies to,

$$P(A_1) = \sum_{k=1}^{\mathcal{M}} P(A_1 \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k) \tag{11.2}$$

The prototypical example appearing in every textbook involves tossing a coin. What is the probability of HEADS on the first toss?

Start off the discussion with just three models. The first model $\mathcal{M}_1$ asserts that the coin is a trick coin with both sides showing HEADS. The second model $\mathcal{M}_2$ asserts that the coin is a "fair" coin with no bias towards either HEADS or TAILS. The third model $\mathcal{M}_3$ asserts that the coin is once again a trick coin, but this time with both sides showing TAILS. A probability of $1/3$ is assigned to each of these three models.

The predictive formula in Equation (11.2) calculates the probability for HEADS on the first trial as,

$$
\begin{aligned}
P(\text{HEADS}) &= P(\text{H}\,|\,\mathcal{M}_1)\,P(\mathcal{M}_1) + P(\text{H}\,|\,\mathcal{M}_2)\,P(\mathcal{M}_2) + P(\text{H}\,|\,\mathcal{M}_3)\,P(\mathcal{M}_3) \\
&= (1 \times 1/3) + (1/2 \times 1/3) + (0 \times 1/3) \\
&= 1/2
\end{aligned}
$$

Now imagine that this symmetry is retained as the number of models is increased. These increasing number of models assign a numerical value for the probability of the trick coin with both HEADS, progressing through to the fair coin, and finally ending up with the trick coin having two TAILS.

Probabilities are assigned across the entire continuum, starting and ending at the trick coins, and for all the biased coins in between, as well as for the fair coin "in the middle." That is, probabilities for HEADS are assigned starting from a probability of 1, marching through the probability of $1/2$, and eventually ending at the probability of 0. In this journey, any legitimate numerical value between 1 and 0 will be assigned.

The probability for each model gets smaller and smaller as the number of models increases, but the probabilities nonetheless remain equal for each model. The coin *could* be a trick coin with two HEADS, it *could* be a biased coin with a favoring of $9/10$ for HEADS, $\cdots$, it *could* be a fair coin, $\cdots$, it *could* be a biased coin with a favoring of $9/10$ for TAILS, it *could* be a trick coin with two TAILS.[1]

The real reason why the probability for HEADS is $1/2$ is because the predictive formula encompasses all such models, not because there are only two alternative outcomes, HEADS or TAILS. The IP knows nothing about the various relative standing of the models, because it does not have possession of any past coin tosses that could serve as data. The value of $1/2$ arises because it is the average of each predicted value between 0 and 1 as made by all of the models.

---

[1] This argument is presented more rigorously in later examples.

11.2.3   Generalizing to many events in the past

With the proof template leading to Equation (11.1), it is not hard to see how to deal with past data in general. Taking the first inductive step, the prediction formula for statement A at time step $N = 3$ based on previous data at $N = 1$ and $N = 2$ would look like,

$$P(A_3 \mid A_1, A_2) = \sum_{k=1}^{\mathcal{M}} P(A_3 \mid \mathcal{M}_k)\, P(\mathcal{M}_k \mid A_1, A_2)$$

If we leap ahead, and say that generally we would like to make a prediction about A at time step $N + 1$ given past data at all N previous time steps, $A_1, A_2, \cdots, A_N$, we write the prediction formula as,

$$P(A_{N+1} \mid \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \mid \mathcal{M}_k)\, P(\mathcal{M}_k \mid \mathcal{D}) \qquad (11.3)$$

where we let $\mathcal{D}$ stand for the sequence of past data $A_1, A_2, \cdots, A_N$.

A numerical example

Consider the situation where the information processor is trying to predict A_3 based on the known outcomes of A_1 and A_2. What state of knowledge will the information processor possess about A_3 after computing the predictive distribution?

Let's stick with the coin tossing scenario. Suppose A_3 is the statement, "HEADS will show on the next toss." A_1 is the statement, "TAILS appeared on the first toss." A_2 is the statement, "HEADS appeared on the second toss." Keep the same $\mathcal{M} = 3$ models under consideration that assert a fair coin and the two trick coins.

At this point, we can fill in all the ingredients for the predictive formula,

$$P(\text{HEADS}_3 \mid \text{TAILS}_1, \text{HEADS}_2) = \sum_{k=1}^{3} P(\text{HEADS}_3 \mid \mathcal{M}_k)\, P(\mathcal{M}_k \mid \text{TAILS}_1, \text{HEADS}_2)$$

Now the summation on the right hand side will reduce to just one term because two of the models will be impossible given the outcome of the first two tosses. The two trick coin hypotheses can be ruled out because, in fact, we observed a TAILS on the first toss ruling out the two–HEADED trick coin, and then a HEADS appeared on the second toss ruling out the two–TAILED trick coin. The only model remaining is the hypothesis that the coin is fair. The formal rules for prediction must always subsume what is deductively clear.

Thus, our state of knowledge about getting HEADS on the next toss is,

$$P(\text{HEADS}_3 \mid \text{TAILS}_1, \text{HEADS}_2) = [\, 1 \times 0\,] + [\, 1/2 \times P(\mathcal{M}_2 \mid \text{TAILS}_1, \text{HEADS}_2)\,] + [\, 0 \times 0\,]$$

We find that we are left with a subproblem to solve: What is the updated probability for the second model given the observed data? This subproblem is a specific example of the more general question: What is the updated probability for any k^{th} model, $P(\mathcal{M}_k \,|\, \mathcal{D})$, after seeing some data?

11.2.4   Updating the state of knowledge about the kth model

If, in fact, there are statements concerning the data known to be true, then the probability assigned to the various models can be updated by conditioning on these known data. Like any probabilistic updating, we make use of Bayes's Theorem.

$$P(\mathcal{M}_k \,|\, \mathcal{D}) \;=\; \frac{P(\mathcal{M}_k, \mathcal{D})}{P(\mathcal{D})}$$

$$=\; \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}{P(\mathcal{D})}$$

$$=\; \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}$$

Completing the numerical example

With this result, we can complete the numerical example started in the last section.

$$P(\mathcal{M}_2 \,|\, \text{TAILS}_1, \text{HEADS}_2) = \frac{P(\text{TAILS}_1, \text{HEADS}_2 \,|\, \mathcal{M}_2)\; P(\mathcal{M}_2)}{\sum_{k=1}^{3} P(\text{TAILS}_1, \text{HEADS}_2 \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}$$

We already know that two of the terms in the denominator are 0. Under the fair coin hypothesis,

$$P(\text{TAILS}_1, \text{HEADS}_2 \,|\, \mathcal{M}_2) = 1/2 \times 1/2 = 1/4$$

The prior probability assigned to this hypothesis was,

$$P(\mathcal{M}_2) = 1/3$$

Thus,

$$P(\mathcal{M}_2 \,|\, \text{TAILS}_1, \text{HEADS}_2) = \frac{1/4 \times 1/3}{0 + (1/4 \times 1/3) + 0} = 1$$

A mechanical application of the formal rules for probabilistic inferencing must always yield a sensible result. Here we have the sensible result that, after deductively eliminating two of the three models, the only remaining model must have an updated probability of 1. It is certain to be the only operative model after the first two coin flips have taken place.

Plugging this result into the predictive formula,

$$P(\text{HEADS}_3 \,|\, \text{TAILS}_1, \text{HEADS}_2) = (1 \times 0) + (1/2 \times 1) + (0 \times 0) = 1/2$$

Again, this seems to be an eminently sensible result. The fair coin model was supported to the maximum extent possible, and the available data are in complete agreement with this revised state of knowledge.

11.3 Data Driven Predictions

Even though the formal paradigm refers, in general, to statements, and is not restricted just to observations, inference in science is usually thought of as being driven by observing vast amounts of data. Our predictions are thought to get better with increasing amounts of experimental observations. As a corollary, bad models are weeded out, and good models should attain the lion's share of attention.

Pressing our humble coin tossing example to the extreme, let's examine these musings with another numerical example of the predictive formula. Suppose now the coin has been flipped 100 times, and it was accurately recorded that HEADS showed 62 times and TAILS 38 times. These observations are the data we are going to leverage to revise our state of knowledge about HEADS appearing on the next flip of the coin.

Once again, for ease of presentation, we will restrict ourselves to just three models. The wisdom of experience suggests that the coin supplied will never be a trick coin if the supplier knows that it will be flipped a number of times, but it still might be one heavily biased in favor of HEADS or TAILS. Therefore, the first model assigns a legitimate numerical value of 3/4 for HEADS to appear. The second model is the benchmark model of a fair coin with a numerical value of 1/2 for HEADS to show. The third and final model is the hypothesis of a coin biased for TAILS, so the numerical value of 1/4 is assigned to HEADS.

Guided only by intuition, in what general direction will our state of knowledge about the next flip move to? The evidence from the data indicates that the third model for a coin biased towards TAILS is discredited, while the other two models, the fair coin and the coin biased towards HEADS, seem to be roughly equally supported. Thus, the probability should be somewhere between 1/2 and 3/4. The quantitative embodiment in the predictive formula had better return an answer in consonance with this intuition.

11.3.1 What about the 101st toss?

The symbol $\mathcal{D}$ is a concise representation for a possible sequence of statements,

$$(A_1 = a_1), (A_2 = a_2), \cdots, (A_{99} = a_2), (A_{100} = a_1)$$

For example, $(A_1 = a_1)$ is the statement, "HEADS appeared on the first flip.";
$(A_2 = a_2)$ the statement, "TAILS appeared on the second flip."; and so on up to
$(A_{100} = a_1)$, the statement, "HEADS appeared on the last flip."

The goal is to capture the state of knowledge about HEADS appearing on the next, that is, the 101st flip of the coin. Thus, we seek $P(A_{101} \,|\, \mathcal{D})$. Use the predictive formula shown as Equation (11.3),

$$P(\text{HEADS}_{101} \,|\, \mathcal{D}) = \sum_{k=1}^{3} P(\text{HEADS}_{101} \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, \mathcal{D})$$

Repeating now the numerical assignments for the probability of HEADS under each of the three models, we have,

$$P(\text{HEADS}_{101} \,|\, \mathcal{M}_1) \;\; = \;\; 3/4$$

$$P(\text{HEADS}_{101} \,|\, \mathcal{M}_2) \;\; = \;\; 1/2$$

$$P(\text{HEADS}_{101} \,|\, \mathcal{M}_3) \;\; = \;\; 1/4$$

The only non–trivial computation resides in Bayes's Theorem where the updated probability for each model must be computed based on the experimental data.

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{k=1}^{3} P(\mathcal{D} \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}$$

Calculate the *likelihood* of the ordered sequence of the data under each model,

$$P(\mathcal{D} \,|\, \mathcal{M}_1) \;\; = \;\; (3/4)^{62} \, (1/4)^{38}$$

$$P(\mathcal{D} \,|\, \mathcal{M}_2) \;\; = \;\; (1/2)^{62} \, (1/2)^{38}$$

$$P(\mathcal{D} \,|\, \mathcal{M}_3) \;\; = \;\; (1/4)^{62} \, (3/4)^{38}$$

The prior probability assigned to each of the three models was,

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = P(\mathcal{M}_3) = 1/3$$

Thus, after calculations based on Bayes's Theorem for the updated probability of each model, we find that,

$$P(\mathcal{M}_1 \,|\, \mathcal{D}) \;\; \approx \;\; 0.2313$$

$$P(\mathcal{M}_2 \,|\, \mathcal{D}) \;\; \approx \;\; 0.7687$$

$$P(\mathcal{M}_3 \,|\, \mathcal{D}) \;\; \approx \;\; 0.0000$$

So, finally we can write out,

$$P(\text{HEADS}_{101} \,|\, \mathcal{D}) = (3/4 \times 0.2313) + (1/2 \times 0.7687) + (1/4 \times 0) = 0.5578$$

We observe that the prediction from the third model, the hypothesis that the coin was biased towards TAILS, is not counted at all. The prediction from the first model, the hypothesis that the coin is biased towards HEADS, and the prediction from the second model, the hypothesis that the coin was fair, are weighted to the extent they were supported by the data. A probability slightly over 1/2 for the next coin flip to turn up HEADS confirms our intuitive grasp of what the answer must look like.

11.4   Predictive Formula with Causal Factors

Every prediction problem in scientific inference is going to include a set of causal factors. These causal factors are either known, or under experimental control, and the model presumes that they influence the event that is to be predicted in some significant way. We amend the formal rules to take account of these causal factors with the consequence that the prediction formula is changed slightly.

Let B_{N+1} be a statement that represents the status of a causal factor. A_{N+1} remains as the statement to be predicted with the presumption, as embodied within some model, that B_{N+1}'s occurrence influences the state of knowledge about A_{N+1}. The whole point is that it is possible to know B_{N+1} at trial $N+1$, but A_{N+1} by its very nature remains uncertain. The past data, $\mathcal{D} \equiv A_1, B_1, A_2, B_2, \cdots, A_N, B_N$, have been jointly observed and accurately recorded.

The same kind of derivation as presented earlier results in the predictive formula,

$$P(A_{N+1} \mid B_{N+1}, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \mid B_{N+1}, \mathcal{M}_k) \, P(\mathcal{M}_k \mid \mathcal{D}) \qquad (11.4)$$

Actually, the second term on the right hand side is also conditioned on the causal factor. However, we would like to keep the same expression for the revised state of knowledge about the models. Therefore, the information about B_{N+1}, and its impact on revising the state of knowledge about the kth model is voluntarily ignored. If there are lots of data, then dropping B_{N+1} won't make much difference anyway on the state of knowledge about the model.

11.4.1   A numerical exercise involving a coin and a causal factor

Here is an even stronger illustration of the data driven enterprise. Suppose that the causal factor B is the stipulation that HEADS is face up and parallel to the table, as the coin now is simply dropped from the short height of three inches as opposed to being flipped in the air.

A graduate student in statistics is given the task of dropping the coin $N = 10,000$ times under these conditions with the causal factor as just described. A is retained as the statement concerning what face was showing after the coin was dropped. The coin showed HEADS 8,104 times and TAILS on the remaining 1,896 drops.

We would like to use a state of knowledge to predict the next drop of the coin, $A_{N+1} = \text{HEADS}$, conditioned on knowledge that the coin was dropped HEADS up three inches from the table, and conditioned as well on the previous data consisting of,

$$(A_1, B_1), (A_2, B_2), \cdots, (A_{9,999}, B_{9,999}), (A_{10,000}, B_{10,000})$$

The three reigning models assert that under the given experimental conditions,

$$P(A \mid B, \mathcal{M}_1) \;=\; 0.70$$

$$P(A \mid B, \mathcal{M}_2) \;=\; 0.80$$

$$P(A \mid B, \mathcal{M}_3) \;=\; 0.90$$

Inserting this information into the predictive formula results in,

$$P(A_{N+1} = \text{HEADS} \mid B_{N+1} = 3 \text{ INCH DROP WITH HEADS UP}, \mathcal{D}) =$$

$$[\, 0.70 \times P(\mathcal{M}_1 \mid \mathcal{D}) \,] + [\, 0.80 \times P(\mathcal{M}_2 \mid \mathcal{D}) \,] + [\, 0.90 \times P(\mathcal{M}_3 \mid \mathcal{D}) \,]$$

The numerical details will be explored in an exercise. Essentially what happens is that $\mathcal{M}_2$ is so overwhelmingly supported by the data that the other two competing models are thoroughly discredited. Model $\mathcal{M}_2$ is left as the sole model standing since,

$$P(\mathcal{M}_2 \mid \mathcal{D}) \approx 1 \text{ and } P(\mathcal{M}_1 \mid \mathcal{D}) \approx 0 = P(\mathcal{M}_3 \mid \mathcal{D}) \approx 0$$

This model asserts a degree of belief equal to 0.80 that HEADS will occur on the next drop given that the coin was dropped from a height of three inches oriented with HEADS UP. This model's prediction will not be averaged together with the predictions of the other two models. Of course, our intuition would have told us that this model was the one most heavily favored by the observations, but we might have been surprised at the decisiveness of the quantitative evidence.

11.5   Connections to the Literature

I want to use this opportunity to stress that the correct answer to these issues was essentially given by both Bayes and Laplace over 200 years ago. Nevertheless, there ensued persistent and unrelenting criticism, seemingly steeped in a profound misunderstanding about their analysis, setting back inferencing for generations. An excellent source on the technical historical details is provided in Anders Hald's [9]

A History of Mathematical Statistics from 1750 to 1930

See especially Chapter 15 about Laplace's **Rule of Succession**.

The correct inference about the *next* event when N trials have already been observed MUST follow the formal manipulation rules of probability. Laplace's **Rule of Succession** *does* follow these rules. It is the proper way to conduct an inference. These formal rules are reflected in the general template of Equation (11.3),

$$P(A_{N+1} \mid \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \mid \mathcal{M}_k)\, P(\mathcal{M}_k \mid \mathcal{D})$$

By these same formal rules (here Bayes's Theorem), the second term on the right hand side of Equation (11.3) becomes,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(\mathcal{D})}$$

Following these rules leads to Laplace's Rule of Succession. Furthermore, Bayes, Laplace, and Jaynes all used a uniform prior for $P(\mathcal{M}_k)$.

We see in $P(\mathcal{M}_k)$ the *prior probability* for the models. It represents the IP's state of knowledge *about the models* prior to any data.

If the IP is "totally ignorant" or "completely uninformed" about the models, then it MUST assign them equal standing, or what is the same thing, a uniform or flat distribution. To do otherwise would be to admit that something *is* known about the relative status of the models. Granting a different status to the models violates the assumption that the IP must begin in a totally uninformed state of knowledge about all the causes for A at this beginning point in the chain of inference.

The most basic application of the **Rule of Succession** was treated in section 11.2.2. There we found that the probability for the very first occurrence of HEADS, prior to knowing anything at all about the coin, or its method of being tossed, was 1/2. This result depended on using equal probabilities for the models under consideration.

This was Bayes's reasoning, this was Laplace's reasoning, and it was repeated by Jaynes in re–deriving Laplace's **Rule of Succession**, It also happens to be my reasoning.

If I may be permitted one final comment. Laplace's language and thought processes were centered on *causes* and *events*. Today, we call them models and statements. Nevertheless, the underlying concept is still the same.

Laplace, in dividing up probability problems, stated that one class of problems involved finding the probabilities for causes given some known events (the origin of the label *inverse probability*). If, he said, at the beginning of the inference, the reasoner had no reason to lend more credence to any one supposed cause over another, he should go ahead and assign equal probabilities to all of the causes. The resulting observations would properly re–order this initial equiprobability about the causes.

His *Principle of Insufficient Reason* was meant to apply to not knowing anything about the *causes*; NOT to not knowing anything about the *events*. Therefore, it *is* correct to say that we arrive at a probability of 1/2 for HEADS by invoking a principle involving "insufficient reason."

This conclusion is reached NOT because we have no reason to favor either of the two statements in the state space, but rather because we have no reason to favor any of the models assigning their probabilities.

This reliance on an "insufficient reason to prioritize causes" naturally leads to the uniform probability distribution for all models. Each model does its job by making a numerical assignment from 0 to 1 to HEADS. In other words, as Laplace surmised, the correct state of knowledge about the next coin toss is obtained only if the reasoner who knows nothing about the causes of the event uses the correct rules of probability together with a uniform distribution over the causes.

What has confused people even more is that the correct application of the rules of probability to future events, (namely, Laplace's **Rule of Succession**), leads to equiprobability for seeing no HEADS, one HEAD, two HEADS, ..., or N HEADS in N future tosses of the coin when nothing is known about the coin, or its manner of being tossed.

This correct result is NOT because the principle of insufficient reason has been invoked either at the level of *statements*, or at the level of N *future events*, but rather because it has been invoked at the level of *causes*. And moreover, this correct inference MUST involve the uniform prior probability for models.

This fundamental concept behind the **Rule of Succession** is misunderstood by almost everybody, including our most renowned mathematicians, probabilists, and statisticians, to include Bayesians and believers in the Maximum Entropy Principle. What *they* do not seem able to comprehend would have been dismissed as a mere platitude by Bayes and Laplace. It is a miracle that we manage to make any progress at all in advancing human knowledge.

11.6 Solved Exercises for Chapter Eleven

Exercise 11.6.1. Construct another numerical example with more models illustrating the prediction formula for the first toss of a coin.

Solution to Exercise 11.6.1

We will use the simplified predictive formula,

$$P(A_1 = \text{HEADS}) = \sum_{k=1}^{\mathcal{M}} P(A_1 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

with $\mathcal{M} = 5$ models each assigning a numerical value for the probability of HEADS. $\mathcal{M}_1$ assigns a value of 0.10, $\mathcal{M}_2$ assigns a value of 0.30, $\mathcal{M}_3$ assigns a value of 0.50, $\mathcal{M}_4$ assigns a value of 0.70, and $\mathcal{M}_5$ assigns a value of 0.90 in order to maintain symmetry around 0.50.

$$\sum_{k=1}^{5} P(A_1 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k) = (0.10 \times 1/5) + (0.30 \times 1/5) + (0.50 \times 1/5) +$$

$$(0.70 \times 1/5) + (0.90 \times 1/5)$$

$$P(A_1 = \text{HEADS}) = 0.50$$

The state of knowledge about the first toss ending up HEADS is reflected in a probability of $1/2$ because the information processor is *maximally uncertain about the models* giving rise to the various numerical assignments.

Please note that this computation for the probability of HEADS takes place before any data have been collected.

Exercise 11.6.2. Generalize the first exercise by including all possible legitimate numerical assignments to a probability for HEADS.

Solution to Exercise 11.6.2

This exercise is a beginning to the promised rigor, mentioned earlier, involving symmetry and the limit process. The transition is from the discrete case of $\mathcal{M}$ models to the continuous case of all models assigning numerical values between 0 and 1. The sum in the prediction formula is replaced by an integral with the limits of 0 to 1. With an obvious abuse, the notation is changed to,

$$q \quad \equiv \quad P(A_1 \,|\, \mathcal{M}_k)$$

$$\mathrm{pdf}\,(q) \quad \equiv \quad P(\mathcal{M}_k)$$

The compensation for the notational sloppiness is the resulting easy integration. We'll try to remember that $\mathrm{pdf}\,(q)$ really means a probability for *the model* which assigns the numerical value of q to A_1.

Thus, the prediction formula now looks like,

$$\int_0^1 q\,\mathrm{pdf}\,(q)\,\mathrm{d}q \equiv \lim_{\mathcal{M}\to\infty} \sum_{k=1}^{\mathcal{M}} P(A_1\,|\,\mathcal{M}_k)\,P(\mathcal{M}_k)$$

The analog to maximal uncertainty about the models in the discrete case is a uniform distribution for $\mathrm{pdf}\,(q)\,\mathrm{d}q$ where the *probability density function* takes on the value 1 all along the q–axis from 0 to 1. The integral simplifies and is easily solved as,

$$\int_0^1 q\,\mathrm{pdf}\,(q)\,\mathrm{d}q = \int_0^1 q\,\mathrm{d}q = 1/2\,q^2\,|_0^1 = 1/2$$

We arrive at the same answer as before. The numerical value of $1/2$ is assigned to $P(A_1)$ not because there are only *two* alternatives, HEADS or TAILS, but because the prediction formula converts lack of information about the models into maximal uncertainty about what the models are predicting.

Exercise 11.6.3. Show in detail the derivation for the predictive formula given in section 11.4.

Solution to Exercise 11.6.3

We'll rearrange the order of the proof as first shown in section 11.2.1 for some variety. In this version, start off with Bayes's Theorem,

$$P(A_{N+1}\,|\,B_{N+1},\mathcal{D}) = \frac{P(A_{N+1},B_{N+1},\mathcal{D})}{P(B_{N+1},\mathcal{D})}$$

Now we are going to express the joint probability in the numerator as a marginal sum over all $\mathcal{M}$ models,

$$P(A_{N+1},B_{N+1},\mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1},\mathcal{M}_k,B_{N+1},\mathcal{D})$$

The right hand side will be decomposed according to the **Product Rule**. The variables were placed in this order so that the final two terms in the **Product Rule** decomposition would be $P(B_{N+1}\,|\,\mathcal{D})\,P(\mathcal{D})$. This way these two terms can be taken out of the summation, converted back to a joint probability by a backward

application of the **Product Rule**, and subsequently used as the denominator in Bayes's Theorem.

$$\sum_{k=1}^{\mathcal{M}} P(A_{N+1}, \mathcal{M}_k, B_{N+1}, \mathcal{D}) =$$

$$\sum_{k=1}^{\mathcal{M}} P(A_{N+1} \,|\, \mathcal{M}_k, B_{N+1}, \mathcal{D}) \; P(\mathcal{M}_k \,|\, B_{N+1}, \mathcal{D}) \; P(B_{N+1} \,|\, \mathcal{D}) \; P(\mathcal{D})$$

Divide both sides by $P(B_{N+1} \,|\, \mathcal{D}) \; P(\mathcal{D}) = P(B_{N+1}, \mathcal{D})$ to yield,

$$\frac{P(A_{N+1}, B_{N+1}, \mathcal{D})}{P(B_{N+1}, \mathcal{D})} = \sum_{k=1}^{M} P(A_{N+1} \,|\, \mathcal{M}_k, B_{N+1}, \mathcal{D}) \; P(\mathcal{M}_k \,|\, B_{N+1}, \mathcal{D})$$

Bayes's Theorem tells us what the left hand side is, and, just as before, drop the conditioning on $\mathcal{D}$ in the first term on the right hand side.

$$P(A_{N+1} \,|\, B_{N+1}, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \,|\, \mathcal{M}_k, B_{N+1}) \; P(\mathcal{M}_k \,|\, B_{N+1}, \mathcal{D})$$

As mentioned earlier, drop the conditioning on B_{N+1} in the second term. Rearrange the ordering for the causal factor and the kth model in the first term purely for aesthetic reasons to finally yield,

$$P(A_{N+1} \,|\, B_{N+1}, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \,|\, B_{N+1}, \mathcal{M}_k) \; P(\mathcal{M}_k \,|\, \mathcal{D})$$

Exercise 11.6.4. Provide the missing numerical details for the example in section 11.4.1.

Solution to Exercise 11.6.4

As always,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k) \; P(\mathcal{M}_k)}{P(\mathcal{D})}$$

where,

$$P(\mathcal{D}) = \sum_{k=1}^{3} P(\mathcal{D} \,|\, \mathcal{M}_k) \; P(\mathcal{M}_k)$$

Solving for the revised state of knowledge about model 2,

$$P(\mathcal{D} \,|\, \mathcal{M}_2) \; P(\mathcal{M}_2) = (0.80^{\,8104} \times 0.20^{\,1896}) \times 1/3 \approx e^{-4859.84}$$

Now this is an enormously small number, but similar calculations for the other two models result in even smaller numbers,

$$P(\mathcal{D} \mid \mathcal{M}_1)\, P(\mathcal{M}_1) = (0.70^{\,8104} \times 0.30^{\,1896}) \times 1/3 \approx e^{-5173.23}$$

and,

$$P(\mathcal{D} \mid \mathcal{M}_3)\, P(\mathcal{M}_3) = (0.90^{\,8104} \times 0.10^{\,1896}) \times 1/3 \approx e^{-5219.54}$$

Thus, for all intents and purposes,

$$P(\mathcal{M}_2 \mid \mathcal{D}) = \frac{e^{-4859.84}}{e^{-5173.23} + e^{-4859.84} + e^{-5219.54}} \approx 1$$

and,

$$P(\mathcal{M}_1 \mid \mathcal{D}) = P(\mathcal{M}_3 \mid \mathcal{D}) \approx 0$$

In the statement of the problem, the probability assigned under model $\mathcal{M}_2$ for observing HEADS given the status of the causal factor was,

$$P(A = \text{HEADS} \mid B = \text{Causal factor operative}, \mathcal{M}_2) = 0.80$$

This means that,

$$P(A \mid B, \mathcal{M}_2) = \frac{P(A, B \mid \mathcal{M}_2)}{P(B \mid \mathcal{M}_2)} = 0.80$$

It was certain that the causal factor $B = T$ was operative at every single trial. Thus, $P(B \mid \mathcal{M}_2) = 1$. The data $\mathcal{D}$ were,

$$(A_1 = \text{HEADS}, B_1 = T), (A_2 = \text{TAILS}, B_2 = T), \cdots, (A_{10,000} = \text{HEADS}, B_{10,000} = T)$$

The probability for the data under model $\mathcal{M}_2$ in the precise order in which they occurred was then something like,

$$P(\mathcal{D} \mid \mathcal{M}_2) = \overbrace{0.80 \times 0.20 \times \cdots \times 0.80}^{10,000 \text{ trials}}$$

If the experiment had sometimes set $B = F$, then the data might have been different. More importantly, a joint probability table for A and B under a different model would take into account the fact that the causal factor could have been $B = T$ or $B = F$. $P(B = T \mid \mathcal{M}_\star)$ would no longer be equal to 1 under this model.

Chapter 12

Extending the Formal Rules for Prediction

12.1   Introduction

The template derived in the last Chapter allowed information processors to update a state of knowledge about future events. It is a very general and very powerful procedure based solely on formal manipulation rules.

It is my intent in this Chapter to build on this introduction and develop ever more practical techniques that information processors can exploit for prediction. Eventually, we will get around to seeing how all of this impacts our assessment of predicting the evolution of probabilistic cellular automata.

We begin by exploiting and expanding upon the prediction template for a single future occurrence for some event. In the last Chapter, we pointed out how the generic prediction formula could be simplified when no previous data existed. As a consequence, all the models considered were on an equal footing.

To keep things grounded, we continue with the coin flipping game. However, instead of merely predicting whether the *next* toss will be a HEADS or TAILS, we would like to update a state of knowledge extending indefinitely far into the future.

We are trying to expand our ability to predict further out into the future, not only out to the first flip of the coin, but out to the second, third, and ultimately the Nth flip. Suppose then that the question before us is: What is the information processor's updated state of knowledge about, say, the next *four* coin flips?

Once again, these are questions that Laplace answered correctly a long time ago. Unfortunately, one must re–invent the wheel from time to time.

12.2 Predicting the Indefinite Future

For the next coin flip, that is, the very first coin flip if there were in fact no previous coin flips, and thus no previous data, we developed the prediction formula,

$$P(A_1) = \sum_{k=1}^{\mathcal{M}} P(A_1 \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)$$

This covered the case of some discrete number $\mathcal{M}$ of models assigning numerical values for $P(A_1)$. In an exercise, we extrapolated to the continuous case where the models covered the whole interval of possible numerical assignments from 0 to 1.

$$P(A_1) = \int_0^1 q \, \mathrm{pdf}\,(q) \, \mathrm{d}q$$

Here q is understood to be just a short form for the numerical assignment made by model $\mathcal{M}_k$ to statement A on the first trial, and $\mathrm{pdf}\,(q)$ the uncertainty surrounding the model that made this assignment.

Now we want to extend this formula to cover the situation of $P(A_1, A_2, \cdots, A_N)$ where for numerical practice we'll let $N = 4$. What is the prediction for obtaining, say, two HEADS and two TAILS in the first four coin flips? This kind of summary statement could implicate many different sequences of the $A_1, A_2, \cdots, A_N$.

The model $\mathcal{M}_k$ will assign some legitimate numerical value Q as a probability attached to HEADS and $(1 - Q)$ for the complementary probability attached to TAILS. The capital Q indicates that some definite numerical assignment has taken place for the generic q. This assignment reflects the information inserted by model $\mathcal{M}_k$. We will discuss the general continuous case where the models will encompass every possible assignment for q covering the interval from 0 to 1.

12.2.1 The multiplicity factor

Suppose we voluntarily discard the information about the particular order in which events occur. Then, there are only five things that can possibly happen when we flip the coin four times, and the particular order of the results is ignored. These five possibilities are: 1) all four HEADS, 2) all four TAILS, 3) three HEADS and one TAIL, 4) three TAILS and one HEAD, and, finally, 5) two HEADS and two TAILS. This last possibility is the one that we are interested in predicting.

We just stipulated that the particular order of the sequence of HEADS and TAILS would be ignored. However, if we do take this into account, then each of these five possibilities might happen in more than one way. Two HEADS and two TAILS can happen in more than one way when we consider all the possible order sequences that eventually lead to this aggregate description.

It is easy to exhaustively list all the possibilities for this scenario when only four flips are contemplated. For example, there is the particular sequence, "HEADS on

the first flip, TAILS on the second flip, HEADS on the third flip, and TAILS on the fourth flip." An exercise will show that there are, in fact, a total of six of these kind of possibilities for obtaining two HEADS and two TAILS.

It is clear that there is only *one* way to obtain all four HEADS, or, for that matter, all four TAILS. There are *four* ways to obtain three HEADS and one TAIL, or three TAILS and one HEAD. These various number of ways that each of the five summary statements can happen is called the *multiplicity factor*. W is traditionally used to indicate the multiplicity factor.[1]

A convenient formula exists for calculating the different number of ways these aggregate descriptions can occur. The multiplicity factor for the $n = 2$ case we are examining here is,

$$W = \frac{N!}{N_1! \, N_2!}$$

where N_1 stands for the number of future HEADS, and N_2 for the number of future TAILS. For general n, the multiplicity factor is,

$$W = \frac{N!}{N_1! \, N_2! \, \cdots \, N_n!}$$

Thus, mechanically applying this formula to determine the number of ways that two HEADS and two TAILS might occur in $N = 4$ flips, is calculated as,

$$W = \frac{4!}{2! \, 2!} = 6$$

12.2.2 Probabilities involving the multiplicity factor

Look at the joint probability for four coin flips as conditioned on some model, $P(A_1, A_2, A_3, A_4 \,|\, \mathcal{M}_k)$. Suppose (A_1, A_2, A_3, A_4) is the joint statement "HEADS on first flip, TAILS on second flip, HEADS on third flip, TAILS on fourth flip." This is one of the six ways that two HEADS and two TAILS can occur. The joint probability of this particular sequence is going to boil down to $Q \times (1 - Q) \times Q \times (1 - Q)$.

Remember that the joint probability concerning these future coin tosses can be decomposed according to the **Product Rule**. However, the probability is not dependent on the results of any previous coin flips. The only dependency is on the model assigning the value Q to HEADS or $1 - Q$ to TAILS. So, rather than having to write out the full **Product Rule** expansion as,

$P(A_1, A_2, A_3, A_4 \,|\, \mathcal{M}_k) =$

$\quad P(A_4 \,|\, A_3, A_2, A_1, \mathcal{M}_k) \, P(A_3 \,|\, A_2, A_1, \mathcal{M}_k) \, P(A_2 \,|\, A_1, \mathcal{M}_k) \, P(A_1 \,|\, \mathcal{M}_k)$

---

[1]The use of W seems to stem from Ludwig Boltzmann and the German word for probability, *Wahrscheinlichkeit.* A literal translation into English is something like, "the degree of resemblance to the truth," which seems particularly apt for capturing the essential nature of a probability.

we are allowed to write the simpler,

$$
\begin{aligned}
P(A_1, A_2, A_3, A_4 \,|\, \mathcal{M}_k) \;\; &= \;\; P(A_4 \,|\, \mathcal{M}_k) \times P(A_3 \,|\, \mathcal{M}_k) \times P(A_2 \,|\, \mathcal{M}_k) \times P(A_1 \,|\, \mathcal{M}_k) \\[2mm]
&= \;\; Q^2 \, (1 - Q)^2
\end{aligned}
$$

But this represents only one of the six ways that two HEADS and two TAILS can arise. We must include the other five ways pointed out to us by the multiplicity factor. Now we have,

$$
P(\text{Two HEADS and two TAILS} \,|\, \mathcal{M}_k) = 6 \times [\, Q^2 \times (1 - Q)^2 \,]
$$

The generalization is straightforward. There are going to be a total of $N = 4$ flips, and if we divide up N into N_1 HEADS and N_2 TAILS, then $N_1 + N_2 = N$. Above, we let $N_1 = 2$ and $N_2 = 2$. Thus, we can write the probability for an aggregate statement more generally as,

$$
P(N_1, N_2 \,|\, \mathcal{M}_k) = W \times [\, Q^{N_1} \, (1 - Q)^{N_2} \,]
$$

12.2.3   The prediction equation for four coin flips

We are faced with the same generalization as before when we set up an integral to handle the space of models that assign legitimate numerical values in the range from 0 to 1. For a discrete model space,

$$
P(N_1, N_2) = \sum_{k=1}^{\mathcal{M}} P(N_1, N_2 \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)
$$

and, after generalizing to the continuous model space,

$$
P(N_1, N_2) = \int_0^1 W \times [\, q^{N_1} \, (1 - q)^{N_2} \,] \, \mathrm{pdf}\,(q) \, \mathrm{d}q
$$

The first step in solving this last formula is easy. We just take the constant value of the multiplicity factor outside of the integral because it does not depend on q.

$$
P(N_1, N_2) = W \int_0^1 [\, q^{N_1} \, (1 - q)^{N_2} \,] \, \mathrm{pdf}\,(q) \, \mathrm{d}q
$$

The next major decision involves what form $\mathrm{pdf}\,(q) \, \mathrm{d}q$ should take. Before, in introducing the simpler predictive equation for just one event, we set up a uniform distribution over the space of models. This represented an initial uncertainty about what model or models were actually operative. We will do the same thing here. The probability distribution that captures uniformity over the space of models assigning values from 0 to 1 is the *beta distribution*.

The next steps in the derivation stem from well–known facts in calculus and statistics. We will simply show the result to wrap up this section, and relegate the details to an exercise. The answer is perhaps surprising. The formal rules for prediction state that,

$$P(N_1, N_2) = \frac{1}{N+1} = \frac{1}{5}$$

for all five possible future frequency counts. All five aggregate possibilities possess the same probability. Surprisingly, the probability for obtaining all four TAILS in the first four flips of the coin is the same probability as obtaining two HEADS and two TAILS, namely 1/5.

12.2.4 The origin of the binomial distribution

At first blush, a probability of 1/5 for each of the five outcomes doesn't seem quite right. Shouldn't an outcome like two HEADS and two TAILS have a higher probability simply because there are more ways for such an event to come about when compared to, say, no HEADS and four TAILS? But the proper application of the formal rules for manipulating probabilities always makes sense after we spend a little time re–educating our intuition.[2]

Take care to note that, in the above discussion, we allowed for the entire gamut of possible numerical assignments q via the space of models pdf (q). The conscious choice of unique values for the parameters in the *beta distribution* will result in an "uninformative" state of affairs for the relative standing of all the models. This will accomplish the desired goal of complete freedom for whatever actual physical effects might be operating when absolutely nothing is known about the makeup of the coin, or how it might be tossed.

Everything from trick coins to heavily biased coins to fair coins to almost fair coins can have their day in court. If, in fact, the coin *is* a trick coin with two TAILS, then no HEADS and four TAILS is to be expected, to say the least. Or, even if the coin is not the TAILS trick coin, maybe it is heavily biased to show TAILS, or the coin tosser has some special skill in this regard. Models that assign numerical values to $(1 - q)$ like 0.99, 0.98, and so on, allow for this scenario. So, once again, given such a numerical assignment to $(1 - q)$, no HEADS and four TAILS doesn't seem that outrageous.

But, on the other hand, maybe the coin was fair, or close to fair, so we can't discount this possibility either. In the end, when we mull it over, it does make sense after all that the five aggregate outcomes should have an equal probability.

One way to lock in this new intuition is to run a simulation where a q is randomly selected from the uniform distribution from 0 to 1. Suppose that at some trial, the first random number is $Q_1 = 0.7231$. Consequently, $Q_2 = 0.2769$. Simulate four flips with this Q_1 and Q_2, so that a second random number less than 0.7231 would

---

[2]A felicitous moralizing phrase borrowed from Jaynes.

simulate a HEADS showing on the first flip with, conversely, a simulated TAILS showing if the random number were greater than 0.7231. Repeat this procedure for all four flips. Then add the outcome, say it was three HEADS and one TAIL, to a running total for each of the five outcomes.

Continue this simulation, say for 10,000 trials, *with a new random number and a new Q_1 and Q_2 at each trial.* You will observe that each of the five outcomes occurs about 2,000 times. Remember that *one* trial in the simulation consists of four tosses of the coin.

Why did we have that initial inclination against the result from the prediction equation? It is because we have been indoctrinated into thinking that the binomial distribution should apply in this case. The knee–jerk reaction is to pull the binomial off the shelf for the initial assignment to $P(N_1, N_2)$. This explains the origin for the prejudice in favor of two HEADS and two TAILS. However, we will see that the prediction equation *does* provide a wonderfully revealing answer to this mystery.

If we happen to have settled on only *one* model, then the prediction equation returns the binomial distribution. Technically, the answer depends upon the Dirac delta function, and its properties as defined by integration.

In short, the delta function is written as $\delta(q - Q)$ to replace $\text{pdf}(q)$. This change captures the state of knowledge that only one specific numerical assignment, namely Q, is to be used in the prediction equation. Because of the so–called "sifting property" of the Dirac delta function within an integral, the prediction equation returns the answer of,

$$P(N_1, N_2) = W \times [\, Q^{N_1} \, (1 - Q)^{N_2} \,]$$

Thus, we see that the binomial distribution stems from the prediction equation when there are no past data, no causal factors, and only *one* model making a definite numerical assignment for a probability of HEADS.

Suppose that the coin *is known* to be a fair coin. Then, the information processor might justifiably propose only one model assigning a numerical value of $Q = 1/2$. Therefore, the probability for two HEADS and two TAILS becomes,

$$P(N_1 = 2, N_2 = 2) = \frac{4!}{2! \, 2!} \, (1/2)^2 \, (1/2)^2 = 3/8$$

as compared to the probability for no HEADS and four TAILS,

$$P(N_1 = 0, N_2 = 4) = \frac{4!}{0! \, 4!} \, (1/2)^0 \, (1/2)^4 = 1/16$$

Because of the multiplicity factor, two HEADS and two TAILS is indeed six times more heavily favored than no HEADS and four TAILS. This is correct if it known that the coin's physical make up does not favor either HEADS or TAILS, and, furthermore, will not be tossed using some special skill.

12.3   Discussing Some Generalizations

Finding a state of knowledge for any number N of coin flips taking place in the future is easy. For example, suppose we want to use the prediction equation to develop a probability distribution for $N = 100$ future coin flips. We now know that there are 101 summary descriptions of what could happen when we don't care about the particular sequential order; that is, everything from "100 HEADS and no TAILS." through "50 HEADS and 50 TAILS." to "No HEADS and 100 TAILS." Each one of these aggregate events has the same probability,

$$P(N_1, N_2) = \frac{1}{N + 1} = \frac{1}{101}$$

Through the auspices of the models, none of which we believed in any more than the others, trick coins to fair coins to biased coins of any type were allowed to assign probabilities to HEADS and TAILS. In this manner, we developed an initial state of knowledge for future frequency counts prior to any data, or specification of causal factors.

In the previous sections, we explored in some detail the specific case involving the future uncertainty of four coin flips that had only two possible outcomes. Or, in other words, we were dealing with $N = 4$ and $n = 2$. It is time to tackle the obvious generalizations.

What about the situation where any number of things, other than just the $n = 2$ case of HEADS or TAILS for the coin flip, could happen during one trial? In addition, we would like to retain the right to consider any number of trials extending arbitrarily far into the future in order not to constrain N to any particular number.

This problem is solved in Exercise 12.6.7 for general N and general n. There we find that,

$$P(N_1, N_2, \cdots, N_n) = \frac{N! \, (n - 1)!}{(N + n - 1)!} \tag{12.1}$$

Let's change the game from coin tossing to rolling a die. Now there are $n = 6$ different spots that might show on the die after it lands. There are, as you might expect, a lot more possible frequency counts when rolling the die as compared to tossing the coin. Consider the statement, "Three ONES, no TWOS, one THREE, two FOURS, no FIVES, and four SIXES in ten rolls of the die." as one example of future frequency counts under this new scenario. Again, looking at the total of ten rolls, we don't pay any attention to the specific roll on which these spots showed up.

Nevertheless, we know exactly how many possible future frequency counts there are from Equation (12.1). If each possible frequency count is assigned the same probability by Equation (12.1), then we can invert the formula to find their total

number. For the future event of $N = 10$ rolls of the die where $n = 6$ possibilities exist at each roll, there are a total of,

$$\frac{(N+n-1)!}{N!\,(n-1)!} = \frac{(10+6-1)!}{10!\,5!} = 3,003$$

summary descriptions, one of which was just presented. They all have the same probability of $P(N_1, N_2, \cdots, N_6) = 1/3003$.

Thus, a different frequency count of say "Ten SIXES in ten rolls" has the same probability as the frequency count mentioned above. The space of models takes account of the peculiar circumstance that maybe all six faces of the die show a SIX.

All of these comments are, of course, predicated on setting up a probability distribution for models such that it reflects an uninformative state of knowledge about the q_i. Or, in other words, we are purposely leveraging a lack of information about the relative standing among all the models assigning the numerical values to the faces of the die.

12.4   Taking Account of Past Data

In this manner, we have managed to exploit the prediction equation for the purpose of constructing a state of knowledge about future frequency counts. By definition, this probability distribution could not rely on any past data because there were no past data.

In Chapter Eleven, we derived an extension to the prediction equation for the more general case involving dependency on past data. When data happen to become available, they too can be easily incorporated into an updated state of knowledge through the prediction equation.

Equation (11.3) was presented as this extension of the prediction equation for the next trial A_{N+1} after data $\mathcal{D} = A_1, A_2, \cdots, A_N$ from N previous trials had been recorded.

$$P(A_{N+1} \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, \mathcal{D})$$

To repeat, we are interested in a summary statement about future frequency counts. These are events like the statement, "Two HEADS and two TAILS occur in the next four coin flips." In contrast to the previous discussion, we now have some additional help in forming a state of knowledge about these future events because we know the results of past coin flips. Perhaps the coin was flipped five times in the past with three HEADS and two TAILS showing up.

To resonate with features of statistical physics, we introduce the two phrases *micro–statements* and *macro–statements*. The multiplicity factor W tells us how many different micro–statements are in a macro–statement. A micro–statement is a

more detailed statement like, "HEADS on first flip, TAILS on second flip, HEADS on third flip, TAILS on fourth flip." There are six of these micro–statements, all different in the detail of their sequential ordering, but all resulting in the same macro–statement, "Two HEADS and two TAILS occur in the next four coin flips.".

Thus, we might first write, in consonance with the notation already used,

$$P(N_1, N_2, \cdots, N_n \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(N_1, N_2, \cdots, N_n \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, \mathcal{D})$$

for the discrete case and,

$$P(N_1, N_2, \cdots, N_n \,|\, \mathcal{D}) = \int \cdot \int_{\sum q_i = 1} P(N_1, N_2, \cdots, N_n \,|\, q_i)\, \mathrm{pdf}\,(q_i \,|\, \mathcal{D})\, \mathrm{d}q_i$$

where the model space is continuous.

It is quite tedious, but unfortunately necessary, to keep harping on a notation that keeps everything clear. First, we retain $N_1, N_2, \cdots, N_n$, but now use them to refer to the frequency counts of any *past data* with $\sum_{i=1}^{n} N_i$ still equaling N. This is to adhere to the way N is traditionally thought of in probability and statistics.

Obviously, we now need a new notation to refer to the *future* frequency counts. To that end, we introduce $M_1, M_2, \cdots, M_n$ to refer to the frequency counts of *future events* with $\sum_{i=1}^{n} M_i = M$.

For example, we now write the first term in the prediction equation as,

$$P(M_1, M_2, \cdots, M_n \,|\, q_i) = W(M)\, q_1^{M_1}\, q_2^{M_2} \cdots q_n^{M_n}$$

where these M_i represent frequency counts for each of the n possibilities on future trials. We show an argument M for the multiplicity factor to disambiguate it from the multiplicity factor based on N. This notation for the future frequency counts is not to be confused with $\mathcal{M}$, the total number of models, or $\mathcal{M}_k$, the kth model assigning some numerical value to the q_i.

Using the following equivalency also seems to result in more transparent and easily grasped formulas,

$$P(M_1, M_2, \cdots, M_n \,|\, \mathcal{D}) \equiv P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n)$$

With the notation on the right, we are seeking to be more explicit about capturing a state of knowledge for a macro–statement that talks about *future* frequency counts, the M_i, when conditioned on knowledge of a macro–statement consisting of *previous* frequency counts, the N_i.

As we did before, we will present the final answer along with some easy numerical examples. The road to this final answer is a rather long and involved journey, but the minutiae in all of their detail are set out in the exercises. The probability of some set of frequency counts to occur in the future, given that data already exist in the form of frequency counts for the same n events, is written as,

$$
\begin{aligned}
P(M_1, M_2, \cdots, M_n \,|\, \mathcal{D}) &\equiv P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) \\
&= C \times \frac{\prod_{i=1}^{n}(M_i + N_i)!}{\prod_{i=1}^{n} M_i!} \tag{12.2}
\end{aligned}
$$

where C is some constant term.

12.4.1 Predicting coin flips based on past data

Let's apply this new prediction formula in Equation (12.2) to the coin flipping scenario. We retain the simple state space consisting of only $n = 2$ possibilities at each trial, namely, HEADS or TAILS. We return to our familiar situation of trying to develop a state of knowledge about two HEADS and two TAILS in the next four flips of the coin. When we didn't have any data from past flips, we found that the probability for all five frequency counts was $P(M_1, M_2) = 1/5$.

Suppose now that we have availed ourselves of the opportunity to flip the coin in question four times, and it turned out that two HEADS and two TAILS resulted. Thus, $N_1 = 2$ and $N_2 = 2$ with $N_1 + N_2 = N = 4$. We are uncertain about the *next* four flips and, in particular, uncertain about obtaining $M_1 = 2$ and $M_2 = 2$, that is, getting another two HEADS and two TAILS in the *next* four tosses. We also have that $M_1 + M_2 = M = 4$.

Uncertainty about the future frequency counts can be captured quantitatively through Equation (12.2),

$$
\begin{aligned}
P(M_1 = 2, M_2 = 2 \,|\, \mathcal{D}) &\equiv P(M_1 = 2, M_2 = 2 \,|\, N_1 = 2, N_2 = 2) \\
&= C \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{M_1! \, M_2!} \tag{12.3}
\end{aligned}
$$

We won't worry about calculating the constant C because it happens to have the same value for all five frequencies, or macro–statements.

What we will do is show the relative weighting for all five possibilities to see if anything is different now that we have some previous data. Table 12.1 at the top of the next page shows the calculations using the prediction formula in Equation (12.3) when applied to all five macro-statements.

The second column illustrates each of the five frequencies with two boxes. The first box on the left stands for HEADS and the second box on the right for TAILS. The number in each box is the future frequency count M_1 and M_2 defining the jth macro–statement. The third column shows the appropriate values for M_1 and M_2 inserted into Equation (12.3). The past data, $N_1 = 2$ and $N_2 = 2$, of course remain the same for all five macro–statements.

Given these observations on the coin, things should change around somehow. Previously, all five possible frequencies were on an equal footing because they had

Table 12.1: *Relative weight for five frequencies involving four future flips of a coin when past observations have been made on the coin.*

Macrostatement	Frequencies	Equation (12.3)	Weight
1	4 0	$C \times (6!\, 2!)/(4!\, 0!)$	60
2	0 4	$C \times (2!\, 6!)/(0!\, 4!)$	60
3	3 1	$C \times (5!\, 3!)/(3!\, 1!)$	120
4	1 3	$C \times (3!\, 5!)/(1!\, 3!)$	120
5	2 2	$C \times (4!\, 4!)/(2!\, 2!)$	144

the same initial probability of 1/5. This situation arose because we had to allow for models of things like trick coins before we had access to any data.

But the observations tell us that the coin cannot be either one of the two trick coins. Furthermore, the models of seriously biased coins seem to be discredited by the results from the four previous flips. The available data are not extensive, but nevertheless one wonders how much these few observations might change our state of knowledge.

The five possible frequency counts are revealed to have abandoned the status of equality they enjoyed when no past data were available. The first frequency count predicting all HEADS, and the second frequency count predicting all TAILS, have the lowest relative standing. The frequency count of three HEADS and one TAIL, and its symmetrical counterpart of one HEAD and three TAILS, have twice the weight of the first two. The largest weight is given to the final frequency count, the macro–statement matching the one that occurred during the data gathering phase. All this quantitative rearrangement seems to make sense.

12.4.2   Does the revision make sense with even more data?

Let's conduct another test to see whether the state of knowledge changes in the way that we might expect. What happens when there are a lot more data? Suppose that, in fact, the coin was subjected to 100 previous flips with the observed data of 50 HEADS and 50 TAILS. Now, $N = 100$ with $N_1 = 50$ and $N_2 = 50$. An information processor still wants a probability distribution to capture its state of knowledge for the next four coin flips, so M remains at 4.

Table 12.2, at the top of the next page, is constructed in accordance with the prediction equation of Equation (12.3). The *Weight* column exhibits the very large numbers arising from the calculation of the factorials.

Table 12.2: *The revision in the state of knowledge for the five macro–states after more extensive data have been collected.*

M_1	M_2	N_1	N_2	Equation (12.3)	Weight	Ratios
4	0	50	50	$C \times (54!\,50!)/(4!\,0!)$	2.93×10^{134}	1.00
0	4	50	50	$C \times (50!\,54!)/(0!\,4!)$	2.93×10^{134}	1.00
3	1	50	50	$C \times (53!\,51!)/(3!\,1!)$	1.11×10^{135}	3.77
1	3	50	50	$C \times (51!\,53!)/(1!\,3!)$	1.11×10^{135}	3.77
2	2	50	50	$C \times (52!\,52!)/(2!\,2!)$	1.63×10^{135}	5.56

Including the constant C in the calculation would bring these numbers down to actual probabilities between 0 and 1, but all we want to illustrate is the relative standing among the five macro–statements after being conditioned on the data. The *Ratios* column provides that service for us.

As before, the state of knowledge about the next four flips is moved in the direction that intuition would prefer. With past data of 50 HEADS and 50 TAILS, the two extreme cases of all HEADS or TAILS in four future flips are now ascribed even less weight. Three HEADS and a TAIL, and its symmetrical counterpart, are judged relatively more likely. And finally, an even split in four future tosses is accorded the highest weight of all.

There is a curious pattern revealed in this final column. If just one model, the fair coin model as described by $\delta(q - 1/2)$, were the only model, then the ratios would be the same as reflected by the multiplicity factors. The ratios would look like 1:1:4:4:6. The data from the 100 previous flips are providing the information processor an indication that the model of a fair coin is very strongly supported.

Of course, no amount of finite data can ever provide irrefutable proof in support of just one model, but a relatively modest investment in seeking out some data has resulted in a significant reordering of the model space. In fact, this modest amount of data has advanced us quite a bit down the road to the same spot as the assertion of the dogmatic model where $Q_1 = Q_2 = 1/2$.

12.4.3   Coin flips with more possibilities

This Chapter concludes with a final illustration of the formula for predicting future events when given some past data. As of yet, we have not advanced to the more practical application of considering a set of causal factors thought to have influenced the outcome of the past data. Not unreasonably, we might expect them to continue to influence future events in the same manner as they influenced the observed past events. The next Chapter begins to address this more practical situation.

We might as well keep playing the well–worn coin flipping game as a prelude to playing with probabilistic cellular automata. Consider an objection that someone might raise to the game as described so far. Such critics might reasonably complain that we have been too restrictive in our assumption that only HEADS or TAILS might happen as a result of a flip.

To meet this objection, let's increase $n = 2$ to $n = 4$ to account for two additional possibilities. Not only might the coin land with HEADS or TAILS showing, but the coin might land standing on its edge (it's an especially thick coin we are using), or it might be lost (it rolls off the table and we don't feel like looking for it).

In a more serious vein, what we have just done is to enlarge the *state space*. The state space is denoted by $(A = a_i)$ where i runs from 1 to n, and a_i stands for the ith possible outcome on any one trial. Thus, $(A = a_1) \equiv$ HEADS, $(A = a_2) \equiv$ TAILS, $(A = a_3) \equiv$ EDGE, and $(A = a_4) \equiv$ LOST.

Writing $P(A_{N+1} = a_3)$ is then the way we express the state of knowledge about the statement that the next coin flip will result in the coin standing on its edge. Much more attention should be directed at thinking about the state space for any non–trivial problem. This is merely a tip of the hat to that kind of effort.

We are going to look at joint statements arbitrarily far into the future, so we write,

$$P(A_{N+1} = a_i, A_{N+2} = a_j, \cdots, A_{N+M} = a_k)$$

as the uncertainty about some micro–statement. This statement might be just one of many belonging to an even higher level macro–statement.

Laying the groundwork for a numerical example

Now on to an example illustrating $n = 4$ possibilities at each flip of the coin. Suppose that we are interested in looking into the future to the extent of ascertaining our state of knowledge about the next *five* coin flips. Thus, we begin by setting $M = 5$.

We would like to pose the question: What is my state of knowledge about obtaining three HEADS and two TAILS in the next five flips? We have specified no lost coins or coins standing on edge. We are going to be provided with some past observations on the very same coin involved in the upcoming five tosses. The information processor can never be certain about any outcome that takes place in the future, but inferential reasoning will tell it the degree of certainty that *is* permitted.

To answer this question, we start by substituting the following values: $M_1 = 3$, $M_2 = 2$, $M_3 = 0$, and $M_4 = 0$, with $M = \sum_{i=1}^{4} M_i = 5$.

Suppose, further, that the coin has been flipped $N = 100$ times in the past with the definite result of 50 HEADS, 48 TAILS, 1 EDGE, and 1 LOST. Thus, $N_1 = 50$, $N_2 = 48$, $N_3 = 1$ and $N_4 = 1$ with, of course, $N = \sum_{i=1}^{4} N_i = 100$.

There are, in fact, a total of 56 possible future frequency counts for this problem. This result comes from substituting M for N in the formula for computing the number of macro–statements. The details are worked out in an exercise. Some possibilities might be, "All five HEADS." or, " Two HEADS, one TAIL, one EDGE, and one LOST." or, "All five LOST."

The state of knowledge for the particular frequency count of "Three HEADS, two TAILS, no EDGE, and no LOST." when conditioned on past data from 100 coin flips is now written as,

$$P(M_1 = 3, M_2 = 2, M_3 = 0, M_4 = 0 \,|\, \mathcal{D}) \equiv P(M_1, M_2, M_3, M_4 \,|\, N_1, N_2, N_3, N_4)$$

When the tedious details of processing the factorials in the prediction formula of Equation (12.2) are worked out, the answer is found to be,

$$P(\text{Three HEADS and two TAILS} \,|\, \mathcal{D}) = 0.2574$$

More details on this problem are provided in Exercises 12.6.25 through 12.6.32.

12.5 Connections to the Literature

My detailed explanations of the prediction formula as carried out in the following Exercises were aided by poring over discussions by Aitchison and Dunsmore [1], Bernardo and Smith [2], Frieden [5], Geisser [7], and Jaynes [13].

But the astounding fact, as already mentioned in the last Chapter, is that all of these results were first given by Laplace over two centuries ago. Once again, it is far easier to read Hald's [9] monumental effort in collecting, translating, and putting into more modern notation Laplace's thoughts on these topics than to refer back to Laplace's original scattered and voluminous writings over half a century. Hald's Chapters 9, 10, 11, and 15 are relevant here.

Hald acknowledges, as does almost every historian, that Laplace independently discovered "Bayes's Theorem" around 1774. Hald, and others, often choose the label of *inverse probability* for what today is usually called "Bayesian statistics." For Laplace the only thing that was "inverse" was that he considered the probability of causes when dependent on events as well as the "direct probability" of events dependent on causes.

In the course of his deliberations, Laplace arrived at his infamous **Rule of Succession**. This rule is the same as the prediction formula for the frequency of any number of future events when conditioned on the frequency of past events. The formulas we derived (given our reliance on modern notation) are exactly the ones that Laplace derived.

Here is a verification of one example of the **Rule of Succession** as presented by Laplace in his 1774 paper using my modern derivation of the prediction formula.

Hald treats this example in Chapter 15. Laplace wants to give an example of a problem illustrating the probability of a cause given an event.

Laplace solves this "inverse" problem by finding the probability of drawing a white ticket (*une billet blanc*) from an urn on the very next draw given that some specified number of white tickets and black tickets (*les billets noirs*) have already been picked. Employing the often confusing notation of those earlier times, Laplace stipulated that *p billets blancs* and *q billets noirs*, for a total of $p + q$ *billets*, had already been drawn from the urn. Laplace said that the probability of drawing *une billet blanc* on the next draw was,

$$\frac{p+1}{p+q+2}$$

This is exactly the same situation as predicting HEADS on the next toss of the coin given that we have already observed some number of HEADS and TAILS on previous tosses. Equation (12.2) must therefore provide the same answer as Laplace.

The state space consists of $n = 2$ statements, "A white ticket was drawn." and "A black ticket was drawn.". Since we are interested only in the next draw from the urn, $M = 1$. The notation for this future draw of the white ticket (as opposed to a black ticket) is $M_1 = 1$ and $M_2 = 0$. The white ticket has already been drawn N_1 times in the past, and the black ticket N_2 times in the past, for a total of N draws, or N data points.

Applying the prediction formula, we have,

$$P(M_1 = 1, M_2 = 0 \,|\, N_1, N_2) \;\; = \;\; C \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{M_1! \, M_2!}$$

$$C \;\; = \;\; \frac{M! \, (N + n - 1)!}{N_1! \, N_2! \, (M + N + n - 1)!}$$

$$= \;\; \frac{1! \, (N + 2 - 1)!}{N_1! \, N_2! \, (1 + N + 2 - 1)!}$$

$$= \;\; \frac{(N + 1)!}{N_1! \, N_2! \, (N + 2)!}$$

$$P(M_1 = 1, M_2 = 0 \,|\, N_1, N_2) \;\; = \;\; \frac{(N + 1)!}{N_1! \, N_2! \, (N + 2)!} \times \frac{(N_1 + 1)! \, N_2!}{1! \, 0!}$$

$$= \;\; \frac{N_1 + 1}{N + 2}$$

This is, as hoped, the same answer as found by Laplace with $p \equiv N_1$ and $p + q \equiv N$.

As will be discussed in the technical details, the only way that one arrives at this prediction formula is by using a uniform distribution over the model space. All of the numerical assignments as made by the models, prior to any observations, must be on an equal footing. No one assignment can be given a favored treatment.

This is why $P(\mathcal{M}_k)$ will be given an operational definition that must reflect a "flat" or "uniform" probability density function. None of this adversely affects the correct updating of all the models contingent on the data, that is, $P(\mathcal{M}_k \,|\, \mathcal{D})$.

Now what we with our modern sensibilities prefer to call a "model" is what Laplace considered to be a "cause." According to Laplace, at the outset all possible causes had to be accorded the same respect. The actual "events," the actually observed white and black tickets drawn from the urn, would re–order the initial status of the "causes," the actual unknown constitution of all the tickets in the urn.

The "principle of insufficient reason" advises one, at the very beginning, not to conclude that the urn must possess an equal number of white and black tickets, or only black tickets, or only white tickets, or any number of white and black tickets. This "insufficient reason," so misunderstood by later generations of commentators, refers to the uninformed state of the reasoner about all of these potential causes of the future events, who must perforce lend equal weight to all of these causes.

So the confusing part is that Laplace really was finding $P(\mathcal{M}_k \,|\, \mathcal{D})$ as a necessary preliminary to the solution of the problem, but not saying so explicitly. It was all implicit in his ensuing mathematics. In all of this, it was clear that this must be the starting point for an uninformed information processor, although, of course, Laplace never used this kind of terminology.

Now I don't know whether Laplace actually spelled out in any detail his opinions as I have expressed them here in my own words. But I can't resist thinking that, if we were to ask him today, his response would be, *"Mais oui, Monsieur! Est-ce que cela n'est pas evident?"*

12.6   Solved Exercises for Chapter Twelve

Exercise 12.6.1. Use the formula for the multiplicity factor to find the number of ways to arrive at three HEADS and two TAILS in five flips of the coin.

Solution to Exercise 12.6.1

The multiplicity factor formula for $n = 2$ occurrences at each flip over a total of $N = 5$ flips is,

$$W = \frac{N!}{N_1!\, N_2!} = \frac{5!}{3!\, 2!} = 10$$

There are ten different ways to obtain three HEADS and two TAILS when we pay attention to the specific order sequence of the coin flips. The macro–statement is, "Three HEADS and two TAILS occur during five flips of the coin.", and there are ten micro–statements that fulfill this condition. Set $N_1 = 3$ to specify the three HEADS and $N_2 = 2$ to specify the two TAILS, where $N_1 + N_2$ must equal $N = 5$.

Exercise 12.6.2. Use the formula for the multiplicity factor in exactly the same way to find the number of ways to arrive at one HEAD and four TAILS in five flips of the coin.

Solution to Exercise 12.6.2

The multiplicity factor formula for the defined $n = 2$ possible occurrences at each flip over a total of $N = 5$ flips is,

$$W = \frac{N!}{N_1!\, N_2!} = \frac{5!}{1!\, 4!} = 5$$

There are five different ways to obtain one HEAD and four TAILS when we pay attention to the specific time sequence of the coin flips. The macro–statement is, "One HEAD and four TAILS occur during five flips of the coin." and there are exactly five ways for this to happen. It is clear that $N_1 = 1$ to specify the frequency count of one HEAD and $N_2 = 4$ to specify the frequency count of four TAILS, where, once again, $N_1 + N_2 = 1 + 4 = N = 5$.

It is easy to imagine how these five different outcomes (micro–statements) could have occurred when we take account of the specific sequences over time. All five of these micro–statements must fulfill the condition that one HEAD and four TAILS will occur during the next five tosses of the coin.

1. HEADS on the first flip, TAILS on the final four flips.
2. HEADS on the second flip, TAILS on the first flip, TAILS on the last three flips.
3. HEADS on the third flip, TAILS on the first two flips and TAILS on the last two flips.
4. HEADS on the fourth flip, TAILS on the first three flips and on the final flip.
5. HEADS on the final flip, TAILS on the first four flips.

Exercise 12.6.3. Explicitly list all six different ways to obtain two HEADS and two TAILS in four flips of the coin.

Solution to Exercise 12.6.3

Table 12.3 lists all six different ways. The HEADS column indicates on which flips the two HEADS occurred, and the TAILS column the complementary result on which flips the two TAILS occurred.

Table 12.3: *The listing of all six different ways that a final result of two HEADS and two TAILS could come about in four flips of the coin.*

Way	HEADS	TAILS
1	1st, 2nd	3rd, 4th
2	1st, 3rd	2nd, 4th
3	1st, 4th	2nd, 3rd
4	2nd, 3rd	1st, 4th
5	2nd, 4th	1st, 3rd
6	3rd, 4th	1st, 2nd

Exercise 12.6.4. Look up the expression wherever you please for the multivariate beta distribution (Dirichlet distribution), and then make the correspondence to the version shown below.

Solution to Exercise 12.6.4

My version of the Dirichlet distribution is,

$$\text{pdf}\,(q_1, q_2, \cdots, q_n) = \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \,\prod_{i=1}^{n} q_i^{\alpha_i - 1}$$

An explanation now follows for all of these symbols starting with $\text{pdf}\,(q_1, q_2, \cdots, q_n)$ on the left hand side of the equation. This denotes the abstract joint probability over the n values of the q_i. These q_is represent the numerical assignment under some model to the n possibilities at each trial. Thus, $\text{pdf}\,(q_1, q_2, \cdots, q_n)$ is a probability density for some given model which assigns a legitimate numerical value to each q_i.

DO NOT MISTAKE THIS NOTATION AS INDICATING A PROBABILITY OF A PROBABILITY! For example, in the coin tossing scenario, $n = 2$ and q_1 might be assigned $3/4$, and q_2 assigned $1/4$ by some model. Another model might assign q_1 as 0.01 and q_2 as 0.99. The notation for the joint probability over the q_i, as mentioned before, really refers to $P(\mathcal{M}_k)$. $\mathcal{M}_k$ is the statement that, "Model k assigns the correct numerical values to $q_1, q_2, \cdots, q_n$."

On the right hand side, focus first on the numerator of the fraction. $\Gamma(\cdots)$ refers to the Gamma function. The Gamma function for positive integers turns out to be quite easy. We will discuss it in a few moments. The argument to the Gamma function is the sum of the values assigned to the n parameters in the Dirichlet distribution, the α_i.

Now shift your attention to the denominator of the fraction. The $\prod$ symbol is the analog to the summation symbol. It indicates a repeated multiplication of the Gamma function with the α_i parameter as its argument.

The final term indicates that each q_i will be raised to the $\alpha_i - 1$ power, and multiplied n times.

Exercise 12.6.5. What value for α_i makes for an easy calculation of the multivariate beta distribution (Dirichlet distribution) ?

Solution to Exercise 12.6.5

If we substitute $\alpha_i = 1$ for all n parameters in the multivariate beta distribution, then $\sum_{i=1}^{n} \alpha_i = n$. The numerator reduces to $\Gamma(n)$. And, as alluded to in the previous exercise, the solution to the Gamma function for a positive integer is not complicated; it is just the factorial, $\Gamma(n) = (n-1)!$.

The denominator is equally easy because it is the product of n terms, each of which has the value of $\Gamma(1) = 0! = 1$. Turning now to $\prod_{i=1}^{n} q_i^{\alpha_i - 1}$ with $\alpha_i = 1$, the product reduces to,

$$q_1^0 \times q_2^0 \times \cdots \times q_n^0 = 1 \times 1 \times \cdots \times 1 = 1$$

The final outcome is simply $(n-1)!$ for what initially appeared as the complicated looking multivariate beta distribution after substituting $\alpha_i = 1$ for the parameters. Remember that we are computing a *probability density function* here.

As a quick and dirty check, recall that we said the probability density function for $\text{pdf}(q, 1-q) \equiv \text{pdf}(q_1, q_2)$ took on the value of $(n-1)! = 1$ over the entire q–axis for the $n = 2$ coin flip scenario. The q–axis, (the x–axis), extends from 0 to 1, and the probability density, (the y–axis), takes on the constant value 1. The total area is then 1 as it must be.

Exercise 12.6.6. Building on the results of the last exercise, what does the remaining part of the prediction equation work out to?

Solution to Exercise 12.6.6

Finding the solution for the general case of n is not much harder than for the specific case of $n = 2$, so we will continue to develop the general solution. What remains

in the prediction equation after the multiplicity factor and the probability over the model space have been taken care of is,

$$\int \cdot \int_{\sum q_i = 1} q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n} \, \mathrm{d}q_i$$

This is a shortcut notation that simplifies writing out the $(n-1)$ integral symbols in the multiple integration. This is a classic integration problem that was solved in the 19th Century by the famous German mathematician Peter Gustav Lejeune Dirichlet, and known in his honor as Dirichlet's integral. Without further ado, the answer is,

$$\int \cdot \int_{\sum q_i = 1} q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n} \, \mathrm{d}q_i = \frac{\prod_{i=1}^{n} \Gamma(N_i + 1)}{\Gamma\left[\sum_{i=1}^{n}(N_i + 1)\right]}$$

Look back to the general definition of the multivariate beta presented in Exercise 12.6.4 to discern a pattern operating here. The integral is the inverse of the fraction appearing in front of the n terms consisting of $q_i^{\alpha_i - 1}$. Since the integration of the multivariate beta probability density function over all legitimate values of the q_i must ultimately equal 1, it is clear that dividing 1 by the fractional part results in its inversion. Also, we added 1 to what appeared as the exponent for each q_i in the fractional expression, so the appearance of $N_i + 1$ in the Gamma functions elicits no particular surprise.

All that remains is to work out the numerator and denominator of the answer to Dirichlet's integral in terms of the appropriate factorial functions. The numerator is,

$$\prod_{i=1}^{n} \Gamma(N_i + 1) \quad = \quad \Gamma(N_1 + 1) \times \Gamma(N_2 + 1) \times \cdots \times \Gamma(N_n + 1)$$

$$= \quad N_1! \, N_2! \cdots N_n!$$

The denominator is,

$$\Gamma\left(\sum_{i=1}^{n}(N_i + 1)\right) \quad = \quad \Gamma\left(\sum_{i=1}^{n} N_i + n\right)$$

$$= \quad \Gamma(N + n)$$

$$= \quad (N + n - 1)!$$

Putting these two results together provides us with final piece of the puzzle we require for extending the formal prediction rule.

$$\int \cdot \int_{\sum q_i = 1} q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n} \, \mathrm{d}q_i = \frac{N_1! \, N_2! \cdots N_n!}{(N + n - 1)!}$$

Exercise 12.6.7. Assemble the pieces of the prediction equation to see what results. Find the answer to the question posed in section 12.2.3 by substituting for the specific case of four flips into the future, and only two possibilities at each flip of the coin.

Solution to Exercise 12.6.7

The original prediction equation for general n looked like,

$$P(N_1, N_2, \cdots, N_n) = \int \cdot \int_{\sum q_i = 1} W \times q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n} \, \mathrm{pdf}(q_1, q_2, \cdots, q_n) \, \mathrm{d}q_i$$

Substituting in the results of the last few exercises,

$$P(N_1, N_2, \cdots, N_n) = \frac{N!}{N_1! \, N_2! \cdots N_n!} \times \frac{N_1! \, N_2! \cdots N_n!}{(N + n - 1)!} \times (n - 1)!$$

$$= \frac{N! \, (n - 1)!}{(N + n - 1)!}$$

Substituting $N = 4$ and $n = 2$ we have for all five possible N_1 and N_2,

$$P(N_1, N_2) = \frac{4! \, 1!}{(4 + 2 - 1)!} = \frac{1}{5}$$

Exercise 12.6.8. Look up an explanation for the Dirac delta function. Interpret it from a probability perspective.

Solution to Exercise 12.6.8

The delta function is introduced by thinking of $\delta(q - Q)$ as concentrating on the value Q where q can range from 0 to 1. It arises as a result of a limiting operation where we might begin by considering a probability density function $\mathrm{pdf}(q) \, \mathrm{d}q$ over q, say, from 1/4 to 3/4. Since this covers an interval of 1/2 on the q–axis, the y ordinate would have a constant value of 2 in order that the overall area of the rectangle thus constructed stays at 1 as it must for any probability distribution.

If we narrow the q interval further, say from 0.49 to 0.51, then the height of the rectangle must be 50 to maintain the area at 1. Continue this limiting process until q is concentrated at 0.50, at which point the ordinate must be infinite. We write $\delta(q - 0.50)$ for the probability density function in this case.

One property of the delta function is expressed in terms of integration, and can be viewed from a probability perspective as,

$$\int_0^1 \delta(q - Q) \, \mathrm{d}q = 1$$

which is merely expressing the usual constraint that the sum over all probability assignments must equal 1.

The "sifting property" of the delta function can be thought of probabilistically. It is the average of some function with respect to the probability density function which is represented by the delta function.

$$E[f(q)] = \int_0^1 f(q)\, \delta(q - Q)\, \mathrm{d}q$$

But if q only ever assumes the value of Q, then the average of $f(q)$ must be $f(Q)$. Finally, if $f(q)$ is,

$$P(N_1, N_2 \,|\, \mathcal{M}_k) = W \times [\, q^{N_1}\, (1 - q)^{N_2}\,] \equiv f(q)$$

then the average of this function with respect to the delta function is,

$$P(N_1, N_2) = \int_0^1 W \times [\, q^{N_1}\, (1 - q)^{N_2}\,]\, \delta(q - Q)\, \mathrm{d}q = W \times [\, Q^{N_1}\, (1 - Q)^{N_2}\,]$$

Exercise 12.6.9. How many different ways can each face show up once in six tosses of the die? How many different ways can all SIXES show up in six tosses of the die?

Solution to Exercise 12.6.9

The multiplicity formula tells us that there are,

$$W = \frac{N!}{N_1!\, N_2!\, \cdots\, N_6!} = \frac{6!}{1!\, 1!\, \cdots\, 1!} = 6! = 720$$

different ways for each face to show up once in six rolls of the die. It also tells us that,

$$W = \frac{N!}{N_1!\, N_2!\, \cdots\, N_6!} = \frac{6!}{0!\, 0!\, \cdots\, 6!} = 1$$

there is only one way in which to obtain all SIXES. That one way, of course, is, "SIX on first roll, SIX on second roll, $\cdots$, SIX on last roll."

Exercise 12.6.10. In how many different ways can one FIVE and five SIXes show up in six tosses of the die?

Solution to Exercise 12.6.10

The multiplicity formula tells us that there are,

$$W = \frac{N!}{N_1!\, N_2!\, \cdots\, N_6!} = \frac{6!}{0!\, 0!\, 0!\, 0!\, 1!\, 5!} = 6$$

six different ways for this aggregate description to occur. These six ways are 1) FIVE on first roll, SIXes on last five rolls, 2) SIX on first roll, FIVE on second roll, SIXes on last four rolls, $\cdots$, 6) SIXes on first five rolls, FIVE on last roll.

Exercise 12.6.11. The following thirteen exercises document the detailed derivation of the prediction equation when past data are present. Begin by writing the prediction equation in a simplified notation.

Solution to Exercise 12.6.11

We begin by trying to make the prediction equation as transparent as possible. Let x stand for the past data, y the future event, and θ a parameter. In my exposition, the parameter is the same as the model.

$$P(y \mid x) = \int_\theta P(y \mid \theta) \, P(\theta \mid x) \, d\theta \qquad (12.4)$$

Exercise 12.6.12. Process the second term under the integral of Equation (12.4) by Bayes's Theorem.

Solution to Exercise 12.6.12

By the formal manipulation rule for probabilities that is Bayes's Theorem,

$$P(\theta \mid x) = \frac{P(x \mid \theta) \, P(\theta)}{\int_\theta P(x \mid \theta) \, P(\theta) \, d\theta}$$

Exercise 12.6.13. Substitute this result found in the previous exercise for $P(\theta \mid x)$ into the prediction equation of Equation (12.4).

Solution to Exercise 12.6.13

$$P(y \mid x) = \frac{\int_\theta P(y \mid \theta) \, P(x \mid \theta) \, P(\theta) \, d\theta}{\int_\theta P(x \mid \theta) \, P(\theta) \, d\theta} \qquad (12.5)$$

Exercise 12.6.14. Identify the correspondence of the denominator in Equation (12.5) with a previous result. Stick to the $n = 2$ case.

Solution to Exercise 12.6.14

Since x refers to the past data, the expression for the denominator in Equation (12.5) above corresponds to a result previously written as,

$$\int_\theta P(x \mid \theta) \, P(\theta) \, d\theta \equiv \int_0^1 P(N_1, N_2 \mid q) \, \mathrm{pdf}\,(q) \, dq = P(N_1, N_2) = \frac{1}{N+1}$$

Exercise 12.6.15. Concentrate now on the three terms in the numerator of Equation (12.5). What is the first term?

Solution to Exercise 12.6.15

The first term is the probability for the future event conditioned on some model assigning a numerical value to q,

$$P(y \mid \theta) \equiv P(M_1, M_2 \mid q) = W(M) \, q_1^{M_1} \, q_2^{M_2}$$

Exercise 12.6.16. What is the second term?

Solution to Exercise 12.6.16

Analogous to the first term,

$$P(x \mid \theta) \equiv P(N_1, N_2 \mid q) = W(N) \, q_1^{N_1} \, q_2^{N_2}$$

Exercise 12.6.17. What is the third term?

Solution to Exercise 12.6.17

This is the probability assigned to the models. Once again, use the probability density function captured by the *beta distribution*,

$$P(\theta) \equiv \mathrm{pdf}(q_1, q_2) = C_{Beta} \times q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1}$$

Exercise 12.6.18. Put all three terms back together under the integral in the numerator. Pull out all the constant factors.

Solution to Exercise 12.6.18

First, putting all three pieces back into the numerator yields,

$$\int_0^1 W(M) \, q_1^{M_1} \, q_2^{M_2} \, W(N) \, q_1^{N_1} \, q_2^{N_2} \, C_{Beta} \, q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1} \, \mathrm{d}q$$

Next, pull out from under the integration symbol all the constant factors that do not depend upon q,

$$W(M) \times W(N) \times C_{Beta} \times \int_0^1 q_1^{M_1} \, q_2^{M_2} \, q_1^{N_1} \, q_2^{N_2} \, q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1} \, \mathrm{d}q$$

Exercise 12.6.19. Gather together all the q_i terms. Let $\alpha_i = 1$ for a uniform distribution over the space of models.

Solution to Exercise 12.6.19

$$W(M) \times W(N) \times C_{Beta} \times \int_0^1 q_1^{M_1+N_1} \, q_2^{M_2+N_2} \, \mathrm{d}q$$

Exercise 12.6.20. Substitute the definition of the Dirichlet integral that appears as the last term.

Solution to Exercise 12.6.20

$$W(M) \times W(N) \times C_{Beta} \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{(M + N + 1)!}$$

Exercise 12.6.21. Substitute the factorial expressions for the first three terms in the above expression.

Solution to Exercise 12.6.21

$$\frac{M!}{M_1! \, M_2!} \times \frac{N!}{N_1! \, N_2!} \times (n-1)! \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{(M + N + 1)!}$$

Exercise 12.6.22. Finish up the derivation of Equation (12.5). Stay with the case where $n = 2$.

Solution to Exercise 12.6.22

We have been concentrating on the expression in the numerator of Equation (12.5) as shown in Exercise 12.6.13. We have arrived at this answer,

$$\int_\theta P(y \,|\, \theta) \, P(x \,|\, \theta) \, P(\theta) \, \mathrm{d}\theta = \frac{M!}{M_1! \, M_2!} \times \frac{N!}{N_1! \, N_2!} \times (n-1)! \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{(M + N + 1)!}$$

In Exercise 12.6.14, the denominator in Equation (12.5) was shown to equal,

$$\int_\theta P(x \,|\, \theta) \, P(\theta) \, \mathrm{d}\theta = \frac{1}{N + 1}$$

Dividing the numerator by the denominator to arrive at our final objective of $P(y \,|\, x)$, the probability for the future frequency counts conditioned on the already observed frequency counts, yields,

$$P(M_1, M_2 \,|\, N_1, N_2) = \frac{\frac{M!}{M_1! \, M_2!} \times \frac{N!}{N_1! \, N_2!} \times (n-1)! \times \frac{(M_1+N_1)! \, (M_2+N_2)!}{(M+N+1)!}}{\frac{1}{N+1}}$$

Multiplying top and bottom by $N + 1$ increases $N!$ in the second term to $(N+1)!$. Since $n = 2$, the third term reduces to 1. As a final answer, we are left with,

$$P(M_1, M_2 \mid \mathcal{D}) = \frac{M!}{M_1! \, M_2!} \times \frac{(N+1)!}{N_1! \, N_2!} \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{(M + N + 1)!}$$

Exercise 12.6.23. Collect all the constant terms into C. Write out the final result as given in section 12.4.1 as Equation (12.3).

Solution to Exercise 12.6.23

The data are known, by definition, so N_1, N_2, $N = N_1 + N_2$, and $N + 1$ are known. M has been specified because that is how far into the future the prediction equation is supposed to cover, so M and $(N+M+1)$ are known as well. Thus, we can collect,

$$C = \frac{M! \, (N+1)!}{N_1! \, N_2! \, (M + N + 1)!}$$

as the constant term. The future frequency counts of M_1 for HEADS, and M_2 for TAILS are, of course, unknown, so any term including these future frequencies is not included in the constant term. That leaves the final form of the prediction equation to be written as,

$$P(M_1, M_2 \mid N_1, N_2) = C \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{M_1! \, M_2!}$$

Exercise 12.6.24. What is the constant term C in Table 12.1 that provides the actual probabilities for each macro–statement?

Solution to Exercise 12.6.24

First, write down the expression for the constant term that we have just derived in the previous exercise,

$$C = \frac{M! \, (N+1)!}{N_1! \, N_2! \, (M + N + 1)!}$$

Substitute the known values from the problem to find C,

$$N = 4$$

$$M = 4$$

$$N_1 = 2$$

$$N_2 = 2$$

$$C = \frac{4! \, 5!}{2! \, 2! \, 9!}$$

$$= \frac{1}{504}$$

$$= 0.001984$$

Table 12.4 lists the five macro–statements from Table 12.1, and the actual predictive probabilities $P(M_1, M_2 \mid \mathcal{D})$ assigned to them based on the above computation of C. The normalization factor could also have been found by adding up the five weights.

Table 12.4: *The same five macro–statements of Table 12.1 referring to future frequency counts of HEADS and TAILS in four tosses of a coin. The probabilities for these as yet unknown frequency counts are computed given that we now know how to compute C.*

Frequencies	C	Weight	Predictive Probability
4 0	0.001984	60	0.1190
0 4	0.001984	60	0.1190
3 1	0.001984	120	0.2381
1 3	0.001984	120	0.2381
2 2	0.001984	144	0.2857
Sums		504	1.0000

Exercise 12.6.25. Write out a joint statement that characterizes one of the micro–statements belonging to the macro–statement discussed as an example in section 12.4.3.

Solution to Exercise 12.6.25

One of the ten micro–statements making up the macro-statement, "Three HEADS and two TAILS in five future tosses." is,

$$(A_{N+1} = a_1, A_{N+2} = a_2, A_{N+3} = a_1, A_{N+4} = a_1, A_{N+5} = a_2)$$

where $(A = a_1) \equiv$ HEADS, $(A = a_2) \equiv$ TAILS, $(A = a_3) \equiv$ EDGE, $(A = a_4) \equiv$ LOST, are the $n = 4$ possibilities at each trial. The three HEADS are postulated to occur on future trials 1, 3, and 4, while the two TAILS are postulated to occur on future trials 2 and 5.

Exercise 12.6.26. How many different micro–statements comprise this one macro–statement?

Solution to Exercise 12.6.26

Use the multiplicity formula to calculate that there are ten different ways that this macro–statement could be formed.

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3! \, M_4!} = \frac{5!}{3! \, 2! \, 0! \, 0!} = 10$$

Exercise 12.6.27. Give an example of another macro–statement.

Solution to Exercise 12.6.27

"Two HEADS, one TAIL, one EDGE, and one LOST occur in five future tosses."

Exercise 12.6.28. How many different micro–statements go to make up this macro–statement?

Solution to Exercise 12.6.28

Use the multiplicity formula to calculate that there are sixty different ways that this macro–statement could be formed,

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3! \, M_4!} = \frac{5!}{2! \, 1! \, 1! \, 1!} = 60$$

The enumeration of all sixty micro–statements might begin,

1. "1st HEADS, 2nd HEADS, 3rd TAILS, 4th EDGE, 5th LOST."

2. "1st HEADS, 2nd TAILS, 3rd HEADS, 4th EDGE, 5th LOST."

3. "1st HEADS, 2nd TAILS, 3rd EDGE, 4th HEADS, 5th LOST."

. . .

60. "1st LOST, 2nd EDGE, 3rd TAILS, 4th HEADS, 5th HEADS."

Exercise 12.6.29. How many total macro–statements are there in the numerical example of section 12.4.3?

Solution to Exercise 12.6.29

There are a total of 56 macro–statements for $M = 5$ and $n = 4$.

$$\text{\textbf{Number of frequency counts}} \quad = \quad \frac{(M + n - 1)!}{M! \, (n - 1)!}$$

$$= \quad \frac{(5 + 4 - 1)!}{5! \, 3!}$$

$$= \quad 56$$

Two of these 56 macro–statements have just been discussed in the previous exercises.

Exercise 12.6.30. Write out the probability for the macro–statement "three HEADS and two TAILS" given a numerical assignment via some model $\mathcal{M}_k$.

Solution to Exercise 12.6.30

This is the first term on the right hand side of the prediction equation. The fact that the model $\mathcal{M}_k$ is given indicates that some legitimate numerical assignment has already been made for the abstract probabilities $q_1, q_2, \cdots, q_n$. The notation for this definite numerical assignment is $Q_1, Q_2, \cdots, Q_n$.

Since there are now four possibilities at each flip of the coin instead of two, $n = 4$. Implicitly, we have specified no occurrences of either a coin landing on its edge or getting lost. Thus, the probability for the macro–statement is going to be the sum over all the ways that three HEADS, two TAILS, no EDGES, and no LOSTS could occur in five flips of the coin. The sum is encapsulated in the multiplicity factor $W(M) = 10$ as calculated above in Exercise 12.6.26.

$$
\begin{aligned}
P(M_1 = 3, M_2 = 2, M_3 = 0, M_4 = 0 \,|\, \mathcal{M}_k) &= W(M) \times [\, Q_1^{M_1} \, Q_2^{M_2} \, Q_3^{M_3} \, Q_4^{M_4} \,] \\
&= 10 \times Q_1^3 \, Q_2^2 \, Q_3^0 \, Q_4^0 \\
&= 10 \times Q_1^3 \, Q_2^2
\end{aligned}
$$

Exercise 12.6.31. We have been discussing a macro-statement involving three HEADS and two TAILS in five future flips of a coin. For which of the following two conditions does the information processor have less certainty about whether this statement is true: 1) considering only the data and all possible models, or 2) one specific model assigning 1/2 as a probability for both HEADS and TAILS?

Solution to Exercise 12.6.31

Under the first condition, $P(M_1, M_2, M_3, M_4 \,|\, \mathcal{D}) = 0.2574$ as calculated in the next exercise. Under the second condition, as begun in the last exercise, we found that,

$$P(M_1, M_2, M_3, M_4 \,|\, \mathcal{M}_k) = W(M) \times [\, Q_1^{M_1} \, Q_2^{M_2} \, Q_3^{M_3} \, Q_4^{M_4} \,]$$

$$= 10 \times (1/2)^3 \, (1/2)^2$$

$$= 0.3125$$

There is less certainty about the truth of the future event of three HEADS and two TAILS in five flips of the coin under the process of averaging over all possible models where things like the coin landing on its edge or getting lost could happen. Comparing this situation with, say, a coin known to be fair which could not land on its edge, and, furthermore, would not be lost, we find that the probability is a bit higher. This is as it should be. But, all in all, the difference in the state of knowledge under the two conditions is not as profound as one might have initially thought.

Exercise 12.6.32. Carry out the details of the computation to determine $P(M_1, M_2, M_3, M_4 \,|\, \mathcal{D})$ in section 12.4.3.

Solution to Exercise 12.6.32

We show Equation (12.2) again, and provide the explicit formula for the constant factor C for dealing with general n.

$$P(M_1, M_2, M_3, M_4 \,|\, \mathcal{D}) \quad \equiv \quad P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n)$$

$$= \quad C \times \frac{\prod_{i=1}^{n}(M_i + N_i)!}{\prod_{i=1}^{n} M_i!}$$

Substituting for the known n, M_i, and N_i of this problem, and momentarily leaving C as it is, we have,

$$P(M_1, M_2, M_3, M_4 \,|\, \mathcal{D}) \quad = \quad C \times \frac{(M_1 + N_1)! \, (M_2 + N_2)! \, (M_3 + N_3)! \, (M_4 + N_4)!}{M_1! \, M_2! \, M_3! \, M_4!}$$

$$= \quad C \times \frac{53! \, 50! \, 1! \, 1!}{3! \, 2! \, 0! \, 0!}$$

The value of C, in a derivation similar to the previous exercises, can be shown to equal,

$$C \quad = \quad \frac{M! \, (N + n - 1)!}{N_1! \, N_2! \, N_3! \, N_4! \, (M + N + n - 1)!}$$

$$= \quad \frac{5! \, 103!}{50! \, 48! \, 1! \, 1! \, 108!}$$

Putting both of these factorial expressions back together, and then performing the cancelations,

$$P(M_1, M_2, M_3, M_4 \,|\, \mathcal{D}) = \frac{5! \, 103!}{50! \, 48! \, 1! \, 1! \, 108!} \times \frac{53! \, 50! \, 1! \, 1!}{3! \, 2! \, 0! \, 0!}$$

$$= \frac{53 \times 52 \times 51 \times 50 \times 49 \times 10}{108 \times 107 \times 106 \times 105 \times 104}$$

$$= 0.2574$$

Exercise 12.6.33. Revisit the problem first presented in section 11.3.1. Consider a continuum of models rather than the restricted set of three models discussed earlier.

Solution to Exercise 12.6.33

We wanted to assess the uncertainty surrounding the next toss of the coin after the coin had already been tossed 100 times with the result of 62 HEADS and 38 TAILS. The known values in the equation can be filled in with,

$$n = 2, M_1 = 1, M_2 = 0, M = 1, N_1 = 62, N_2 = 38, \text{ and } N = 100$$

The probability of the future event HEADS on the next toss when conditioned on the data of the 100 past observations is,

$$P(M_1 = 1, M_2 = 0 \mid N_1 = 62, N_2 = 38)$$

Using the prediction equation developed in Equation (12.3),

$$P(M_1, M_2 \mid \mathcal{D}) = C \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{M_1! \, M_2!}$$

$$C = \frac{M! \, (N + 1)!}{N_1! \, N_2! \, (M + N + 1)!}$$

$$= \frac{1! \, 101!}{62! \, 38! \, 102!}$$

$$P(M_1, M_2 \mid \mathcal{D}) = \frac{1! \, 101!}{62! \, 38! \, 102!} \times \frac{(1 + 62)! \, (0 + 38)!}{1! \, 0!}$$

$$= \frac{63}{102}$$

$$= \frac{N_1 + 1}{N + 2}$$

$$P(\text{HEADS on next flip} \mid \mathcal{D}) = 0.6176$$

The former result of 0.5578 may have seemed a little off–putting, but the important fact to remember is that only three models were used. The many potential models assigning a numerical value greater than 0.75 were never considered. Thus, their contributions to the prediction were never taken into account. Now that all models assigning values between 0 and 1 have been incorporated into the prediction, we see a result that is quite in line with our intuition, a probability close to the frequency count of the past coin tosses.

Exercise 12.6.34. Review the derivation of the basic expression for the probability of a first event viewed as some future event from fundamental manipulation principles.

Solution to Exercise 12.6.34

Here is a standard three–step template used in many derivations.

1. Form a marginal sum from the joint probability of the statement and the space of models.

2. Apply the **Product Rule** to the joint probability.

3. Substitute the **Product Rule** decomposition back into the marginal sum.

Let (A_1) stand for some statement considered both as the very first trial and as a future frequency count. Then, $M = 1$ and $N = 0$,

$$P(A_{N+M}) \;=\; \sum_{k=1}^{\mathcal{M}} P(A_1, \mathcal{M}_k)$$

$$P(A_1, \mathcal{M}_k) \;=\; P(A_1 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

$$P(A_1) \;=\; \sum_{k=1}^{\mathcal{M}} P(A_1 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

The notation of the statement $(A_{N+M}) \equiv (A_1)$ clearly indicates that no previous data exist, and that this is the very first (future) trial.

Exercise 12.6.35. What is the generalization from this template as given in the last exercise to the extension of M future trials?

Solution to Exercise 12.6.35

Let $(A_1, A_2, \cdots, A_M)$ stand for some statement considered both as the very first M trials and as a future frequency count. N is still 0, and we write,

$$P(A_1, A_2, \cdots, A_M) \;=\; \sum_{k=1}^{\mathcal{M}} P(A_1, A_2, \cdots, A_M, \mathcal{M}_k)$$

$$P(A_1, A_2, \cdots, A_M, \mathcal{M}_k) \;=\; P(A_1 \,|\, \mathcal{M}_k) \times P(A_2 \,|\, \mathcal{M}_k) \times \cdots \times P(A_M \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

$$P(A_1, A_2, \cdots, A_M) \;=\; \sum_{k=1}^{\mathcal{M}} q_1^{M_1}\, q_2^{M_2} \cdots q_n^{M_n}\, P(\mathcal{M}_k)$$

There are n possibilities at each trial and M_i is the number of times the ith possibility occurred over a total of M trials.

The second term is a probability for the model $\mathcal{M}_k$ that assigns the numerical values $q_1, q_2, \cdots, q_n$,

$$P(\mathcal{M}_k) \equiv \text{pdf}\,(q_1, q_2, \cdots, q_n)$$

Finally, the summation over the space of $\mathcal{M}$ models is a generic representation for a multiple integral covering all potential assignments for the q_i.

$$\sum_{k=1}^{\mathcal{M}} \cdots \equiv \int \cdots \int_{\sum q_i = 1} \mathrm{d}q_i$$

This notation is also supposed to account for the constraint that the total probability assigned to the n possibilities at each trial must equal 1. Thus, for an $n = 3$ case, if q_1 happens to have been assigned a value of 0.90, then q_2 can only be chosen such that $q_2 \leq 0.1$. And, if q_2 is assigned 0.07 following this restriction, there is no freedom of choice for q_3 which must be assigned 0.03.

Strictly speaking, this means that some authors write,

$$\int \cdot \int_{\sum q_i = 1} \text{pdf}\,(q_1, q_2, 1 - (q_1 + q_2))\, \mathrm{d}q_1\, \mathrm{d}q_2$$

whereas aesthetically I prefer instead, somewhat incorrectly,

$$\int \cdot \int_{\sum q_i = 1} \text{pdf}\,(q_1, q_2, q_3)\, \mathrm{d}q_1\, \mathrm{d}q_2\, \mathrm{d}q_3$$

The exercises employ this more transparent notation. The equivalency shown below is also present,

$$\int_\theta P(y\,|\,\theta)\, P(\theta)\, \mathrm{d}\theta \equiv \int \cdot \int_{\sum q_i = 1} P(F_j\,|\,q_1, q_2, \cdots, q_n)\, \text{pdf}\,(q_1, q_2, \cdots, q_n)\, \mathrm{d}q_i$$

where F_j refers to a jth future macro–statement.

Exercise 12.6.36. An information processor is completely uninformed. If something has happened once, what is the probability that it will happen again on the very next trial?

Solution to Exercise 12.6.36

Our general formula for future frequency counts takes notice of all past data, so it will answer that question for us.

$$P(M_1, M_2, \cdots, M_n\,|\,\mathcal{D}) = C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

Something has happened once, so let $N_j = 1$ with all other $N_i = 0$. Therefore, $N = 1$. What is unknown, and what we are uncertain about, is whether that same event will occur on the next trial.

Thus, $M_j = 1$ with all other $M_i = 0$ and $M = 1$ as well. The constant term works out to,

$$
\begin{aligned}
C &= \frac{M!\,(N+n-1)!}{N_1!\,N_2!\cdots N_n!\,(M+N+n-1)!} \\[2ex]
&= \frac{1!\,(1+n-1)!}{0!\,0!\cdots 1!\cdots 0!\,(1+1+n-1)!} \\[2ex]
&= \frac{n!}{(n+1)!} \\[2ex]
&= \frac{1}{n+1}
\end{aligned}
$$

The non–constant term works out to,

$$
\begin{aligned}
\frac{\prod_{i=1}^{n}(M_i+N_i)!}{\prod_{i=1}^{n} M_i!} &= \frac{0!\,0!\cdots 2!\cdots 0!}{0!\,0!\cdots 1!\cdots 0!} \\[2ex]
&= 2
\end{aligned}
$$

Putting these two results back together yields the answer for the fully written out conditional probability of,

$$
P(M_1 = 0, M_2 = 0, \cdots, M_j = 1, \cdots, M_n = 0 \,|\, N_1 = 0, N_2 = 0, \cdots, N_j = 1, \cdots, N_n = 0)
$$

as,

$$
P(M_1, M_2, \cdots, M_n \,|\, \mathcal{D}) = \frac{2}{n+1}
$$

For example, if an IP is totally ignorant about the physical characteristics of a coin, as well as the manner in which it will be tossed, having observed that TAILS appeared on the first flip, the degree of belief that the following statement is TRUE, "TAILS will also appear on the next flip." is accorded a value of 2/3 along the scale from 0 to 1.[3]

Likewise, if an IP is totally ignorant about the physical characteristics of a die, as well as the manner in which it will be rolled, having observed that FOUR appeared on the first roll, the degree of belief that the following statement is TRUE, "FOUR will also appear on the next roll." has a probability of 2/7.

---

[3] This is the answer that caused so much heartburn for Sir Harold Jeffreys when applied to situations like observing just once the formation of water from Hydrogen and Oxygen.

Chapter 13

Predicting College Success

13.1   Introduction

We now have at our disposal some general and powerful formulas for predicting
future events. These formulas were derived by strictly following the formal rules for
manipulating probabilities. Within the conceptual division we have set up, these
kind of rules are important because they are true regardless of how we think about
how numerical values might be assigned to probabilities.

Laplace also followed these manipulation rules, and in consequence, derived his
infamous **Rule of Succession**.[1] We think it helpful to examine the consequences
of this Rule in some detail. Why is it important to do this?

Ever since Laplace first presented his arguments in 1774, unceasing debate about
its meaning and applicability has filled the literature. Some of this debate reveals
a gross misunderstanding, some inadvertent, but for the most part quite willful.

Despite the fact that the Rule has been around for about two and a half centuries,
it is fascinating in the extreme to observe how it has been subsequently mangled
and mis–interpreted. If someone presents a fallacious counter–argument to bury
the **Rule of Succession**, can we trust him elsewhere? Is there not a seed of
doubt planted somewhere in your mind when you turn to his explanations on other
foundational issues?

In any case, close examination of Laplace's line of reasoning will serve to further
illustrate my own development of the foundational principles underlying Information
Processing. The reader may judge for himself which style of reasoning is more to
his liking.

---

[1]This was not Laplace's appellation. Subsequent writers, with a somewhat pejorative intent,
gave it this label.

As our inferential problem, we are interested in predicting future events when bereft of any observations. Let us put on the back burner, temporarily, trying to predict the future when we have the advantage of some known data. We must first understand this simpler situation when data are lacking before assessing how known observations will impact our state of knowledge. Even for this simpler situation, mass confusion reigns supreme.

Let us recall the lessons for inferences about coin tosses in previous Chapters. It is not too hard to then extrapolate to inferences about dice rolls. With this as a warm up, we make the transition to predicting events for a more pragmatic real–world problem: predicting college success for any number of future students.

We have used the turn of phrase "future events." It turns out that we cannot use the word *event* in a loose, cavalier fashion. It has acquired a technical meaning within probability. Along the way to discovering how phrases like *simple events* and *compound events* must be used, we will discuss the equally important foundational concepts of how events are related to the *state space* and the *sample space.*

It soon becomes clear that, with the tools at our disposal, we can directly address the question of how many students will come to share particular patterns, here characterized by graduation and score status. Or, even more abstractly, we learn to deal with how things in general, when subjected to repeated trials, might be distributed with regard to a set of traits.

As mentioned, we begin by restricting the discussion to predicting future joint characterizations of students without the benefit of any previous data. This is the same problem we wrestled with when trying to predict whether the next toss of a coin was HEADS or TAILS when we knew absolutely nothing about the physical make–up of the coin, or the skill, or lack thereof, of the person making the toss.

From the perspective of the formal manipulation rules, the solution depends upon how we choose to create models, and the uncertainty surrounding the model space that is generated.

When there are absolutely no data, it is easy to descend into medieval theological debates. *How many angels can dance on the head of a pin?* Any prediction you might care to make has an equal claim to validity!

However, not being in possession of any data is the least of our worries. Every statement such as "My first name is David and it will rain today." must eventually be judged as TRUE or FALSE. An information processor attaches a probability to such statements indicating a state of knowledge reflecting the information inserted by some model as to the truth or falsity of this statement. **If only in principle**, a correct judgment can always be ultimately rendered. This occurs when the requisite measurements concerning my name and the weather become available.

A statement like "An angel can dance on the head of a pin." can never be adjudicated because the requisite measurements will never become available (as far as we now suspect). Therefore, it is a statement outside the purview of inference.

A measurement or observation in the form of definitely judged outcomes on angels, dancing or not dancing, will never be made. The debate will continue, but it will not be settled through an inference where the original model space reflecting any initial ignorance can be modified through data.

That is why it is helpful to consider examples like coin tosses, rolls of the die, or students graduating. Statements in these scenarios *can* be adjudicated correctly by an IP. Whatever inferential problem we might be involved with, it is a prerequisite that, in principle, an unequivocal judgment can be made as to whether a statement is TRUE or FALSE.

It is quite instructive when attempting to disentangle the conceptual confusion that pervades the distinction between *frequencies* and *probabilities*, to introduce the upcoming elementary counting arguments. The combinatorial formulas developed for simply counting up the ways things could happen shed considerable light on the interplay between frequencies, probabilities, and models. Further examination of the Dirichlet distribution shows how important it is in understanding the debate surrounding "complete ignorance."

13.2 Coins to Dice

In the last Chapter, we relied on the concrete example of coin tosses to advance our argument, or should we say, to advance Laplace's argument? We found out that our prediction of future tosses depended critically on the models chosen to assign numerical values to the probabilities for the two statements in the state space, as well as on the relative importance of each of these models.

Every discussion on probability eventually gets around to rolling the dice after tiring of coin tosses. This transition affords us the chance to bump up the dimension of the state space from $n = 2$ to $n = 6$.

But more important than increasing the size of the state space is an opportunity to see what happens when Laplace's "probability of causes" enters the fray. What happens if we deviate from Bayes's and Laplace's recommendation for equality of causes? To that end, let's discuss rolling two dice from the probabilistic perspective as developed in the last couple of Chapters.

Pose the standard question: What is the probability of seeing the same face after two dice are rolled? Conversely, what is the probability of seeing different faces after the dice are rolled? Inevitably, the answer given is that the probability for the same number of spots on the dice is 1/6, and, conversely, 5/6 for different spots.

But is this correct? What can we say about such an answer when we use our prediction formulas? What role does our knowledge or ignorance about the models play in arriving at an inference about these future events?

13.2.1   State space and sample space

Rolling two dice at the same time is the same as rolling one die twice in succession. Suppose that the two dice are colored red and green. In rolling one die twice, we may choose to disregard the temporal order in which the faces appeared; in rolling two dice at the same time, we may choose to disregard the color of the dice on which the faces appeared.

If we ask about the probability of seeing a TWO and SIX spot, we implicitly acknowledge that we don't care whether the SIX or the TWO came up on the first roll or the second roll. If we roll two differently colored dice at the same time, we don't care whether the TWO shows up on the red die or the green die. This is another example of voluntarily discarding information.

In any case, the interest resides in predicting (an inference) the occurrence of two future outcomes of the dice roll. Therefore, we can bring in our prediction formula that has $P(M_1, M_2, \cdots, M_6)$ on the left hand side for calculating our state of knowledge about any number of future frequency counts. We are dealing here with an especially simple case where $M = 2$.

The state space consists of the six statements, "A ONE spot appeared uppermost after the die was rolled.", "A TWO spot appeared uppermost after the die was rolled.", $\cdots$,"A SIX spot appeared uppermost after the die was rolled." Each one of these statements can only be TRUE or FALSE, and each die roll must have one, and only one, of these statements TRUE. Therefore, the dimension of the state space is $n = 6$.

It is very important to distinguish between the *state space* and the *sample space*. The state space, of course, lists all n possibilities that could be observed, as we have defined the universe of possible measurements, on any roll of one die. The sample space takes account of the number M, the total number of future events the IP is concerned about. We use the language that the sample space consists of *elementary points*. The number of elementary points in the sample space is n^M.

For our dice rolling example then, the total number of elementary points in the sample space will be $n^M = 6^2 = 36$. These elementary points reflect the most refined reduction in terms of a statement that we can make in the discussion of our inferential problem. But except for the simplest of problems, it can be *too* refined.

13.2.2   Aggregating elementary points

We usually are much more interested in macro–statements, or *events* as they are typically called. These statements are formed by aggregating some number of the lowest level elementary points. As we mentioned, we can't get more detailed than listing any of these elementary points. One of the 36 elementary points might be, "SIX on the red die and TWO on the green die.", or, "FIVE on the first roll and ONE on the second roll."

Consider the question introduced at the beginning about whether both dice show the same number of spots, or whether they show different numbers. There are six elementary points that go to make up the higher level statement that both dice show the same number of spots. These are, "Both dice show a ONE spot.", "Both dice show a TWO spot.", $\cdots$, "Both dice show a SIX spot.".

There are thirty elementary points available to fifteen statements that go to make up the higher level statement asserting that the dice show different spots. These are, "The dice show a ONE spot and a TWO spot.", "The dice show a ONE spot and a THREE spot.", $\cdots$, the dice show a FIVE spot and a SIX spot.".

At this juncture, it is not too onerous to exhaustively list these situations in order to double–check that they are correct. But we would very much like to have some convenient combinatorial formulas for counting. Let's begin to take a look at these formulas that alleviate the mental effort in counting up statements occurring at various levels.

13.2.3   Convenient formulas for counting

We've determined that there are six lower level statements for the way we want to aggregate the points to form the first macro–statement. There are fifteen lower level statements that comprise the second macro–statement. This gives us a total of twenty one such statements. These statements are, in fact, the 21 possible frequency counts, or contingency tables, for two future rolls of the dice.

These contingency tables can be rendered graphically as, say, | 0 | 0 | 2 | 0 | 0 | 0 | for observing a THREE spot on both dice, and | 1 | 0 | 0 | 0 | 0 | 1 | for observing a ONE spot and a SIX spot on the dice. The first contingency table is one of the six possible contingency tables where the dice show the same number of spots, while the second contingency table is one of the fifteen contingency tables that show different spots on the dice.

Each one of the six contingency tables where the dice show the same number of spots can happen in only one way. A THREE must have occurred on the green die as well as the red die when two dice were rolled, or a THREE occurred on the first roll and a THREE occurred on the second roll when one die was rolled twice in succession.

However, there are two ways that each of the fifteen contingency tables can come about. The ONE spot occurred on the green die and the SIX spot on the red die, or the ONE spot occurred on the red die and the SIX spot occurred on the green die if we are thinking that two differently colored dice were rolled at the same time. If one die were rolled twice in succession, then the SIX spot could have occurred on the first roll and the ONE spot on the second roll, or the ONE spot could have occurred on the first roll and the SIX spot on the second roll.

This is simply the multiplicity factor. There is only one way that the same number of spots can occur on the two dice. There are two ways that different spots can occur on the two dice.

So we see how the total number of elementary points are aggregated into these kind of higher level statements. There are a total of 36 elementary points in the sample space. The first set of six contingency tables comprise $6 \times 1 = 6$ elementary points, and the second set of fifteen contingency tables comprise the remaining $15 \times 2 = 30$ elementary points. We see that the multiplicity factor entered as the number of ways that the frequency counts could have arisen.

Here then is the formula for computing the total number of contingency tables, or frequency counts, for our dice rolling example,

$$\frac{(M + n - 1)!}{M! \, (n - 1)!} = \frac{(2 + 6 - 1)!}{2! \, 5!} = 21$$

As mentioned, these 21 statements are divided into the 6 statements where the same spot comes up on both dice, and the 15 statements where the dice have different spots.

There is another counting formula which we will use to compute the number of contingency tables for each of the two macro–statements,

$$\frac{n!}{r_z! \times r_s! \times r_d!} = \frac{6!}{5! \, 0! \, 1!} = 6$$

The n in the numerator refers to the dimension of the state space. The symbols in the denominator refer to the *repetition* of zero frequency counts (r_z), the repetition of single frequency counts (r_s), and the repetition of double frequency counts (r_d) in the contingency table. Think of a contingency table with "0"s in five cells, no "1" frequency counts, and a single "2" in the remaining cell as *frequency counts*.

This formula will correctly compute, as well, the fifteen different configurations for the contingency table not showing the same spots,

$$\frac{n!}{r_z! \times r_s! \times r_d!} = \frac{6!}{4! \, 2! \, 0!} = 15$$

Think of a contingency table with "0"s in four cells, two "1"s in the remaining two cells, and obviously no "2"s as *frequency counts*.

13.2.4   Are counts the same as probability?

With this kind of counting analysis, there are those who arrive at an *incomplete*, and more generally, an *incorrect* inference regarding the dice. There are 6 elementary points from the total of 36 that satisfy the first macro–statement. It is then claimed that the aggregation of these elementary points over the total number of elementary points constitute the probability of an event A, or $P(A) = 1/6$.

There are 30 elementary points from the total of 36 that satisfy the second complementary macro–statement. The aggregation of these elementary points then constitute the probability for the complementary event $\overline{A}$, or $P(\overline{A}) = 5/6$. It is clear that all of these elementary points must have been assigned a probability of $1/36$.

13.2.5 What does the prediction formula say?

We can apply our general prediction formula to this problem to see whether it finds the same answer. An example of the first kind of macro–statement from the total of six such statements is "Both dice show the SIX spot." An example of the second kind of macro–statement from the total of fifteen such statements is, "One die shows a FIVE spot and the other die shows a TWO spot."

There are a total of two future frequency counts, so $M = 2$. For example, $M_3 = 1$ refers to a future frequency count of one THREE, $M_4 = 2$ refers to a future frequency count of two FOURs, and $M_6 = 0$ refers to a future frequency count of no SIXes. There are no previous rolls of the die, so all of the N_i are equal to zero.

Thus, the probability for one of the constituents of the first event, say, the probability for observing two SIXes in the first two rolls of the dice, is written as,

$$P(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 0, M_5 = 0, M_6 = 2)$$

The probability for one of the constituents of the second event, say, the probability for observing a TWO and a FIVE in the first two rolls of the dice, is written as,

$$P(M_1 = 0, M_2 = 1, M_3 = 0, M_4 = 0, M_5 = 1, M_6 = 0)$$

The probability for these specific counts, or for that matter, any possible frequency count or contingency table is found to be equal to,

$$P(M_1, M_2, \cdots, M_6) = \frac{M! \, (n-1)!}{(M+n-1)!} = \frac{2! \, 5!}{7!} = \frac{1}{21}$$

Thus, contrary to the above counting argument, our prediction formula says that $P(A) = 6/21$ and $P(\overline{A}) = 15/21$. The details of these kind of computations are provided in the exercises.

13.2.6 Completely uninformed

Now, the most striking feature of this answer is the fact that it is just the inverse for the counting formula giving the possible number of frequency counts or possible contingency tables. There were a total of 21 different possible frequency counts, and each one of these has the same probability.

But the initial befuddlement is the same as experienced before when dealing with coin tosses. There were five possible frequency counts when a coin was going

to be tossed four times, and each one had the same probability. Why should seeing two HEADS in four future coin flips have the same probability as no HEADS, one HEAD, three HEADS, and four HEADS?

Thinking now about the dice, shouldn't an event that can happen in two different ways, say, the future observation of the TWO and the FIVE, be more probable than something that can happen in only one way, say, the future observation of the two SIXes?

The resolution to this puzzle concerning the dice is the same as it was for the coins. The formal manipulation rules of probability demand this answer under the stated condition that the IP is *completely uninformed*.

This is the same answer as found by Laplace when he was contemplating how to treat the probability of causes. When there was *insufficient reason* to believe in the credibility of any one cause for the dice to show a particular face over any other cause, then the multiplicity factor loses its cachet.

We choose to re–phrase Laplace's explanation involving causes into the language of models. No one model assigning numerical values to the probability for the six faces of the die to show was favored over any other model.

It could be that the dice are really "trick dice" analogous to "trick coins." "Trick dice" could be manufactured such that all the faces have a SIX spot painted on them, in which case the numerical assignment under the model would be $Q_6 = 1$ and the other $Q_i = 0$. The dice might be manufactured "fairly," but the person rolling them has a special skill such that the ONE spot has an assigned numerical value of $Q_1 = 0.25$ with the other Q_i splitting the remaining 0.75.

These two models, as well as every other conceivable assignment, have to be considered as possessing the same relative status. This is so because the IP is *completely uninformed* about what is *causing* the dice to behave the way they do. And, if the IP is completely uninformed, then the state of knowledge about event A is indeed 6/21, not 1/6.

13.2.7   Not uninformed

There is a very enlightening explanation for all of this. And it follows directly from the prediction formula as we have derived it. The probability of event A can be 1/6, but only under one very specific condition. That specific condition is, in some sense, the polar opposite of being completely *uninformed*.

If the IP, rather than being "totally ignorant" about the possible causes for why the dice behave as they do, is instead quite certain that, say, the dice are "fair," then the state of knowledge that the two dice show the same spots is indeed 1/6.

Only *one* model making *one* numerical assignment is considered when it is known that the dice are fair. As a consequence, the full panoply of models making all

conceivable numerical assignments can be ignored in this case. It seems that the IP considers itself comparatively well–informed about the various physical causes when the die is rolled. The IP is able to rule out all manner of "loaded" dice, as well as any "special skill" in rolling the dice.

The IP has gone as far it can go in the opposite direction from Laplace's dictum about the probability of causes. It now asserts that the *only* conceivable cause for the dice to behave as they do is because they are perfect cubes subject to no physical influences that would favor one face over another. The IP, it might be said, now believes it has a *sufficient reason* for one cause, and one cause only.

Thus, under this "fair model," the probability of the future event, "Two THREE spots on the next two tosses" is,

$$P(M_1, M_2, \cdots, M_6) = \frac{W(M)}{n^M} = \frac{1}{6^2} = \frac{1}{36}$$

and probability of the future event, "One FIVE and one TWO on the next two tosses" is,

$$P(M_1, M_2, \cdots, M_6) = \frac{W(M)}{n^M} = \frac{2}{6^2} = \frac{2}{36}$$

where $W(M)$ in the numerator refers to the multiplicity factor, and the denominator is the size of the sample space. Now the number of ways something can happen *is* the deciding factor. All of the elementary points in the sample space have the same probability of $1/n^M = 1/36$.

13.3   Probability of Causes

These almost conflicting suppositions about the probability of what causes the die face to show a particular number naturally leads to wondering whether the problem can be treated in full generality. Both of these situations just discussed, as well as any other supposition about the probability of the causes, lead to different inferences about future frequency counts for the dice. Some of these inferences are quite remarkable. They have an interesting pedigree in the history of probability.

We have managed to ferret out a prediction formula from the generality of the formal manipulation rules. We don't use the language of "causality" quite as bluntly as Laplace. We prefer to talk instead about models making numerical assignments to the probabilities for statements in the state space. However squeamish the modern world might be about causality, determinism, and design, the conceptual notion embedded within $P(\mathcal{M}_k)$ is still roughly the same.

The general template of the prediction formula left us with the prior probability,

$$P(\mathcal{M}_k) \equiv \text{pdf}\,(q_1, q_2, \cdots, q_n)$$

Operationally, using the Dirichlet distribution for the probability density function results in nice analytical solutions. What happens to the probability of the event,

"Both dice show the same spots" under different assumptions for the parameters of the Dirichlet distribution?

The early stages in applying the formal rules to find the state of knowledge about a future event gives us,

$$P(M_1, M_2, \cdots, M_6) = \int \cdot \int_{\sum q_i = 1} W(M) \, q_1^{M_1} \times \cdots \times q_6^{M_6} \, \mathrm{pdf}\,(q_1, \cdots, q_6) \, \mathrm{d}q_i$$

Substitute the Dirichlet distribution for the probability of a model,

$$\mathrm{pdf}\,(q_1, \cdots, q_6) = \frac{\Gamma\left(\sum_{i=1}^6 \alpha_i\right)}{\prod_{i=1}^6 \Gamma(\alpha_i)} \prod_{i=1}^6 q_i^{\alpha_i - 1}$$

Take the two terms not involving the q_i outside of the integral, and show only the right hand side with the understanding that left hand side (not shown) is the probability for some future frequency count $P(M_1, M_2, \cdots, M_6)$,

$$\frac{M!}{M_1! \cdots M_6!} \frac{\Gamma\left(\sum_{i=1}^6 \alpha_i\right)}{\prod_{i=1}^6 \Gamma(\alpha_i)} \int \cdot \int_{\sum q_i = 1} q_1^{M_1} \times \cdots \times q_6^{M_6} \times q_1^{\alpha_1 - 1} \times \cdots \times q_6^{\alpha_6 - 1} \, \mathrm{d}q_i$$

Collect the q_i together after multiplying,

$$\frac{M!}{M_1! \cdots M_6!} \frac{\Gamma\left(\sum_{i=1}^6 \alpha_i\right)}{\prod_{i=1}^6 \Gamma(\alpha_i)} \int \cdot \int_{\sum q_i = 1} q_1^{M_1 + \alpha_1 - 1} \times \cdots \times q_6^{M_6 + \alpha_6 - 1} \, \mathrm{d}q_i$$

The solution of the Dirichlet integral is,

$$\int \cdot \int_{\sum q_i = 1} q_1^{M_1 + \alpha_1 - 1} \times \cdots \times q_6^{M_6 + \alpha_6 - 1} \, \mathrm{d}q_i = \frac{\prod_{i=1}^6 \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^6 (M_i + \alpha_i))}$$

leaving us with,

$$\frac{M!}{M_1! \cdots M_6!} \times \frac{\Gamma\left(\sum_{i=1}^6 \alpha_i\right)}{\prod_{i=1}^6 \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^6 \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^6 (M_i + \alpha_i))}$$

Let $\sum_{i=1}^6 \alpha_i = \mathcal{A}$ and $\sum_{i=1}^6 M_i = M$, and then substitute into the above expression,

$$\frac{M!}{M_1! \cdots M_6!} \times \frac{\Gamma\left(\mathcal{A}\right)}{\prod_{i=1}^6 \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^6 \Gamma(M_i + \alpha_i)}{\Gamma(M + \mathcal{A})}$$

Explicitly show the expansion of the products,

$$\frac{M!}{M_1! \cdots M_6!} \frac{\Gamma\left(\mathcal{A}\right)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2) \cdots \Gamma(\alpha_6)} \frac{\Gamma(M_1 + \alpha_1)\,\Gamma(M_2 + \alpha_2) \cdots \Gamma(M_6 + \alpha_6)}{\Gamma(M + \mathcal{A})}$$

Finally, do some rearranging of the terms to yield,

$$\frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \frac{\Gamma(M_1 + \alpha_1)}{M_1!\,\Gamma(\alpha_1)} \times \frac{\Gamma(M_2 + \alpha_2)}{M_2!\,\Gamma(\alpha_2)} \times \cdots \times \frac{\Gamma(M_6 + \alpha_6)}{M_6!\,\Gamma(\alpha_6)}$$

Use the product notation to compactly represent our final version,

$$P(M_1, M_2, \cdots, M_6) = \frac{M! \, \Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \prod_{i=1}^{6} \frac{\Gamma(M_i + \alpha_i)}{M_i! \, \Gamma(\alpha_i)} \qquad (13.1)$$

In this form, a *Mathematica* program can be written to solve for any future frequency counts for any value of the α_i parameters.

For example, use Equation (13.1) to examine the behavior of the probability for the event we have been discussing in this Chapter, "Both dice show the same spots." When all $\alpha_i = 1$, we already know the answer. This is the "Bayes–Laplace uniform prior" representing total ignorance on the part of the IP about the causes. The probability for two ONEs is,

$$P(M_1 = 2, M_2 = 0, \cdots, M_6 = 0) = \frac{M! \, (n-1)!}{(M+n-1)!} = \frac{2! \, 5!}{7!} = \frac{1}{21}$$

The probability for two TWOs, two THREEs, and so on, is the same and so,

$$P(A) = 6/21 = 0.2857$$

When the models all have the same relative standing as specified by $\alpha_i = 1$, the second counting term, the multiplicity factor $W(M)$ is negated. Only the first counting term counting up the different contingency tables has an effect. There are 21 distinct frequency counts all having the same probability.

By way of counter–point, if we don't fix the α_i at 1, but let all the α_i tend to ∞, then we reach the other anchor point defined by the fair model where all $Q_i = 1/n$. Here the multiplicity factor reigns supreme. As we have already discussed, each elementary point in the sample space has a probability of $1/36$. Then, the fifteen contingency tables that can happen in two ways *do* have a total probability of $30/36$, leading to $P(\overline{A}) = 5/6$.

The really bizarre stuff starts to happen when we let the α_i parameters of the Dirichlet distribution depart from the uniform distribution of $\alpha_i = 1$, and march off in the opposite direction. What happens when the α_i start marching towards zero?

If we let all $\alpha_i = 1/2$, then we are using what some claim to represent "complete ignorance." In this case, applying Equation (13.1) results in,

$$P(A) = 6/16 = 0.3750$$

The probability that both dice show the same spots has increased from 0.2857 to 0.3750. In fact, this characterization of the probability for all the models making numerical assignments was advocated by Sir Harold Jeffreys. It is still prominently used today.

Jeffreys experienced some displeasure at the Bayes–Laplace uniform prior that lent equal weight to all conceivable models. It was felt that having seen something

happen once, under presumed strict causality, implies that repetitions should have a higher probability. Jeffreys's weighting for models accomplishes this objective.

Stop at another way–station on the journey to letting the $\alpha_i \to 0$. Let the $\alpha_i = 1/4$. Now the formula returns an answer of $P(A) = 6/12 = 0.50$. The pattern is clear. The probability for both dice to show the same spots is gradually increasing, or, more generally, if something has happened once, then with increasing probability it should happen again.

Finally, let the $\alpha_i = 0.00001$. Now the formula returns an answer of $P(A) = 0.9995$. We are approaching a probability of $1/6$ for each one of the six low–level macro–statements, "Both dice show a ONE spot.", *etc.*, and therefore approaching a probability of 1 that the higher–level statement is true. By summing over these six events, it is almost certain that both dice show the same number of spots.

This was in fact how the British scientist J.B.S. Haldane conceived of complete ignorance. Things are so strongly linked that an event having happened once, it *must* happen again. The IP's "total ignorance" centers not around this perfect linkage, but rather on not knowing whether it's the ONE spot, the TWO spot, $\cdots$, or the SIX spot that will be linked.

One cannot resist a closing comment. We have just seen the rather surprising consequence of letting the $\alpha_i \to 0$. Now, it's one thing if the same die has been rolled twice in succession, and we say that the second roll must show the same spots as the first roll. After all, it might be one of those trick dice, or the thrower of the die is especially skillful in some way or another. You might expect that whatever "bias" in built into the die will manifest itself again on subsequent rolls.

However, the argument must apply as well to the situation when two dice are rolled at the same time. The mysterious linkage dictates that they must both show the same spots. This argument also applies when the rolling of the two dice is separated by distance and time. The first die is rolled in Atlanta on Tuesday and the second die is rolled in Zurich on Thursday. The two dice must still show the same spots.

In quantum physics, this is an example of "quantum entanglement" introduced by Einstein in the "EPR paradox." It is said to represent a violation of both causality and locality. Quantum events are said to be *non–local* and *non–causal*, and this counter–intuitive notion is captured by letting the $\alpha_i \to 0$ in the model assignment process.

13.4   Dice to College Students

For the final example of this Chapter, let's make that transition promised in the **Introduction**. It's a transition from the artifical world of coins and dice to a more pragmatic real–world situation. Can we make inferences about whether college students will graduate if their performance on a test is known?

Given the tenor of the discussion so far, it will be no surprise when we say that the emphasis will be placed on looking at the probability of the causes for a successful student graduation. What is the impact on assuming different things about the probability of the models making the numerical assignments to joint statements?

As a matter of fact, we don't ratchet up the level of difficulty with this new problem, at least in terms of the state space. The dimension of the state space is actually lowered from $n = 6$ in the dice scenario to $n = 4$ for the college students. But we compensate for this reduction in the size of the state space by asking about a larger number of future events. We now inquire about the future status of $M = 4$ students as opposed to $M = 2$ future rolls of the die.

The size of the sample space increases accordingly. Instead of a sample space consisting of $n^M = 6^2 = 36$ elementary points that we dealt with in the dice problem, we now are faced with a sample space consisting of $n^M = 4^4 = 256$ elementary points.

For the dice problem, one elementary point, that is, the most refined and reduced description possible, was something like, "A THREE occurred when the red die was rolled and a FOUR occurred when the green die was rolled." For this new example, an elementary point might be this one, "Alex scored high on the test and graduated, Beth scored high on the test and did not graduate, Carl scored low on the test and graduated, while Dawn scored low on the test and did not graduate."

"Distinguishability" might refer to the order in time, the color of the die, or now, the name of the student. But, as before, when we voluntarily discard information about distinguishability, we can aggregate these elementary points inhabiting the sample space into larger entities. With only two future frequency counts in the dice rolling scenario, we could decompose this sum into only two classes $2 + 0$, or $1 + 1$. The contingency tables could fall into only two classes, one where both dice showed the same face, or one where they showed different faces.

In the new problem, we can decompose the four future frequency counts, $M = 4$, into these five sums,

1. $4 + 0 + 0 + 0$,

2. $3 + 1 + 0 + 0$,

3. $2 + 2 + 0 + 0$,

4. $2 + 1 + 1 + 0$,

5. $1 + 1 + 1 + 1$.

The sum 4+0+0+0 represents those cases where all four students are placed into the same cell of the contingency table. We could call this situation as one where all of the students share the same "trait," trait here meaning the particular combination

of test score and graduation status. This contingency table, $\boxed{0}\boxed{0}\boxed{4}\boxed{0}$, where all four students share the same trait of graduating with low test scores, is an example.

Likewise, $3 + 1 + 0 + 0$ refers to all those frequency counts where three students share the same trait, and the remaining student has a different trait. This particular contingency table, $\boxed{0}\boxed{3}\boxed{0}\boxed{1}$, where three students share the same trait of not graduating with high test scores, and the final student does not graduate with a low test score, is an example. And so on.

Within each one of these five classes of frequency counts, there are a number of specific ways each can happen. The $4 + 0 + 0 + 0$ pattern can happen in 4 ways, the $3 + 1 + 0 + 0$ pattern in 12 ways, the $2 + 2 + 0 + 0$ pattern in 6 ways, the $2 + 1 + 1 + 0$ pattern in 12 ways, and the $1 + 1 + 1 + 1$ pattern in only one way. Add these up to find that there are 35 different possible frequency counts, or contingency tables. Do not forget that we don't care which particular students show up in each of the four cells for each one of these 35 possible frequency counts,

Now this number of 35 is what we would find when we use our counting formula,

$$\frac{(M + n - 1)!}{M! \, (n - 1)!} = \frac{(4 + 4 - 1)!}{4! \, 3!} = 35$$

The combinatorial formula,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times r_t! \times r_q!}$$

tells us how many specific ways there are within each one of the five patterns. The multiplicity factor,

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3! \, M_4!}$$

tells us how many ways there are when we do take notice of the distinguishability feature, namely that each student has a different name.

We can go no further because we have now reached the level of the elementary points. The counting involved in this descent down the hierarchy must, of course, result in a grand total of 256, the total number of elementary points in the sample space. All of this is made clear in the detailed exercise of the next section.

13.4.1 The counting exercise

Prepare a sketch showing all 35 possible contingency tables for the four future students. Show in detail how the overall sum of 256 elementary points is reached. To accomplish this, Figure 13.1, shown at the top of the next page, summarizes the counting arguments as presented earlier. The boxes show what characteristics the next $M = 4$ students might possess. These are the cells of the contingency table strung out in a single line.
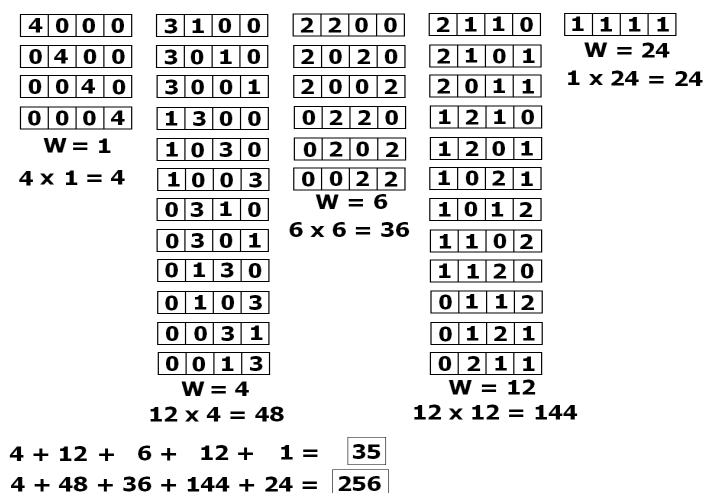
Figure 13.1: *All five possible sums, together with all 35 possible frequency counts (contingency tables), and their associated multiplicity factors as used in the counting exercise.*

The numbers in the boxes reflect the M_i, so they must sum to $M = 4$. The multiplicity factor $W(M)$ for each set of contingency tables is indicated at the bottom of each column. The number of distinct contingency tables in each column is also given as the first term that multiplies $W(M)$.

35 different contingency tables can be assembled from the total of 256 elementary points in the sample space. For example, one frequency count, "Two students graduate with low scores and two students do not graduate with high scores." can happen in six different ways when considering the unique identity of each student.

Prepare a sketch showing how the four students, Alex, Beth, Carl, and Dawn, could be distributed in 24 distinct ways for the statement reflecting the fact that each student has a different trait. Let our four students be abbreviated as **a**, **b**, **c**, and **d**. Figure 13.2 at the top of the next page shows how the multiplicity factor,

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3! \, M_4!} = \frac{4!}{1! \, 1! \, 1! \, 1!} = 24$$

is arrived at for the 35th and final contingency table which asserts that each student is in a different cell, $\boxed{1 \mid 1 \mid 1 \mid 1}$.

There are six rows and four columns. Alex is placed in the first cell in the first column, Beth is placed in the first cell in the second column, and so on. Each one of these 24 possibilities represents an elementary point in the overall sample space of $n^M = 4^4 = 256$ points. Thus, the event that all four students possess a different trait has a probability of 24/256 under the fair model.

	1	**2**	**3**	**4**
1	a b c d	b a c d	c a b d	d a b c
2	a b d c	b a d c	c a d b	d a c b
3	a c b d	b c a d	c b a d	d b a c
4	a c d b	b c d a	c b d a	d b c a
5	a d b c	b d a c	c d a b	d c a b
6	a d c b	b d c a	c d b a	d c b a

Figure 13.2: *All 24 possible ways for four students to arrange themselves one to a cell.*

13.5   Surprising Predictions

What is the IP's state of knowledge that everyone shares the same trait? At first blush, it wouldn't seem that the collection of contingency tables shown as $\boxed{4\,|\,0\,|\,0\,|\,0}$, $\boxed{0\,|\,4\,|\,0\,|\,0}$, $\boxed{0\,|\,0\,|\,4\,|\,0}$, and $\boxed{0\,|\,0\,|\,0\,|\,4}$ where all the students have the same graduation–test score characteristics, could ever become a very probable. After all, they only represent four elementary points in the sample space.

Surprisingly, however, such a probability can indeed approach arbitrarily close to certainty. How can a counter–intuitive assertion like this be true? It implies that the multiplicity factor is not merely being negated as happened when all possible models were considered, but, on the contrary, the situations with the *minimum* multiplicity factor of $W(M) = 1$ are elevated to near certainty.

The answer lies in the generality of $P(\mathcal{M}_k) \equiv \mathrm{pdf}\,(q_1, q_2, \cdots, q_n)$. The Dirichlet distribution has n parameters α_i. These were all set to the $\alpha_i = 1$ "uninformative" model space where all models were placed on an equal footing. However, if these parameters are allowed to vary, the most mysterious things begin to happen.

For example, and as an explanation for our little mystery, if the α_i are all allowed to approach 0 from their initial starting point of 1, then the remarkable thing that ensues is that the probability of the four frequency counts $\boxed{4\,|\,0\,|\,0\,|\,0}$ through $\boxed{0\,|\,0\,|\,0\,|\,4}$ share nearly all the available probability, leaving practically none for the other 31 contingency tables. Thus, each of these four frequency counts approach a probability of 1/4, leaving the other frequency counts with the barest table scraps! The probability that all students share the same trait approaches certainty!

I have already alluded to the historical fact that both Bayes and Laplace asserted that a uniform probability should be used when the IP was "totally ignorant" concerning the "probability of the causes." This notion has proven to be a major

conceptual stumbling block almost from the time it was announced. I will mention only two of the most prominent objectors in this long debate, the iconic figure (at least to Bayesians) of the British geophysicist Sir Harold Jeffreys, and his lesser known fellow British scientist, John Burdon Sanderson Haldane.

13.6   How Laplace Reasoned

To end this Chapter, it seems appropriate to give Laplace the last word. Is there a correspondence between one of our prediction formulas, and Laplace's original argument (1774), later seen as the simplest example of the **Rule of Succession**? Here is the problem as Laplace posed it, dealt with earlier in section 12.5.

> If an urn contains an infinite number of white and black tickets in an unknown ratio, and if one were to draw $p + q$ tickets of which p are white and q are black; what is the probability that in drawing a new ticket from the urn it would be white?

It is clearer if we present the modern argument first, but modify it by following Laplace's line of reasoning. The correspondence between Laplace's notation and ours will be pointed out where necessary. Contrary to the discussion in this Chapter, the inference depends on having seen some previous data, or, as Laplace put it, we have already drawn some number of white and black tickets from the urn.

Equation (11.3) is the prediction formula we derived for the very next occurrence of some statement given that we are in possession of some previously observed data. This is exactly what we require in order to predict that the very next ticket drawn will be a white one conditioned on the known facts that some number of tickets have already been drawn, and their color noted.

We will write Equation (11.3) as,

$$P(A_{N+1} = w \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} = w \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, \mathcal{D})$$

for the probability of the statement that the very next ticket drawn is a white one when conditioned on the known data symbolized by $\mathcal{D}$.

We begin with the second term on the right hand side, $P(\mathcal{M}_k \,|\, \mathcal{D})$. By Bayes's Theorem, we know that this is,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)}$$

The data $\mathcal{D}$ consist of N_1 white tickets and N_2 black tickets for a total of $N = N_1 + N_2$ tickets drawn. The probability for these observations, assuming that the correct numerical assignment has been made under model $\mathcal{M}_k$ is,

$$P(\mathcal{D} \,|\, \mathcal{M}_k) = W(N) \, q_1^{N_1} \, q_2^{N_2}$$

The numerator then becomes,

$$P(\mathcal{D} \,|\, \mathcal{M}_k) = W(N) \, q_1^{N_1} \, q_2^{N_2} \, P(\mathcal{M}_k)$$

The sum in the denominator becomes an integral when we let the q_i take on all possible numerical assignments between 0 and 1 inclusive. The region of integration then must be from 0 to 1. We now have,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{W(N) \, q_1^{N_1} \, q_2^{N_2} \, \text{pdf}\,(q_1, q_2)}{\int_0^1 W(N) \, q_1^{N_1} \, q_2^{N_2} \, \text{pdf}\,(q_1, q_2) \, \mathrm{d}q}$$

The multiplicity factor $W(N)$, which doesn't depend on q, comes outside the integral and cancels.

The very important next step that Laplace took is his invocation of a flat prior for the models. This means that the density function for every single model has the same value. Thus, $P(\mathcal{M}_k)$ and $P(\mathcal{M}_j)$ cancel as well, leaving us with,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{q_1^{N_1} \, q_2^{N_2}}{\int_0^1 q_1^{N_1} \, q_2^{N_2} \, \mathrm{d}q}$$

At this juncture, we can transition back to the way Laplace wrote it. He says that the probability that x is the true ratio of white to black tickets is,

$$\frac{x^p (1 - x)^q \, \mathrm{d}x}{\int x^p (1 - x)^q \, \mathrm{d}x}$$

He explicitly mentions the fact that x will be integrated from 0 to 1. Now, all of the relevant notational translations can be set down.

$$x \equiv q_1$$

$$(1 - x) \equiv q_2$$

$$p \equiv N_1$$

$$q \equiv N_2$$

$$p + q \equiv N$$

What we call a model making a particular numerical assignment of probabilities to statements in the state space, Laplace thought of as "one cause," or "a sufficient reason" for seeing the data. If x were assumed to be the true ratio of white tickets to black tickets in the urn, then this was "sufficient reason" to write an expression like $x^p (1 - x)^q$.

However, if no one sufficient reason existed, or all of the causes, that is, the true ratios, had to be considered as equally plausible, then all of these causes had to be

assigned the same equal probability. Laplace wanted to integrate from 0 to 1 to cover any conceivable true ratio of white to black tickets as $p + q$, the total number of tickets, got larger and larger, (remember he specified that the urn contained an "infinite number" of tickets.) In our conceptualization of the integration, we wanted to make sure that every legitimate assignment of a numerical value to a probability was covered.

In following Laplace's reasoning, we see that he employed Bayes's Theorem. His denominator is,

$$P(\mathcal{D}) \equiv \sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \mid \mathcal{M}_k) \, P(\mathcal{M}_k) \equiv \int x^p (1-x)^q \, \mathrm{d}x$$

But, of course, in 1774 he didn't know it under that label. He had mentioned it earlier in his paper as a fundamental principle, and when invoked, it was, "*par le principe de l'article précédent.*"

With this much accomplished, we can turn our attention to the first term in Equation (11.3). This is the probability that, after the N tickets already drawn from the urn, the very next draw, the $(N + 1)st$ draw, will be a white ticket when conditioned on some model. Well, of course, for Laplace this is simply x, the true ratio of a white ticket on a draw. He now has the expression,

$$\frac{x^{p+1}(1-x)^q \, \mathrm{d}x}{\int x^p (1-x)^q \, \mathrm{d}x}$$

and for us,

$$\frac{q_1^{N_1+1} \, q_2^{N_2}}{\int_0^1 q_1^{N_1} \, q_2^{N_2} \, \mathrm{d}q}$$

The final step, as the prediction formula tells us, is to sum this product over all considered models. Since we want to sum over all legitimate assignments, turn the sum into another integration where q_1 takes on all values from 0 to 1. As mentioned earlier, even though we explicitly show both q_1 and q_2, the integration is a single integration where only q need vary.

$$P(A_{N+1} = w \mid \mathcal{D}) = \frac{\int_0^1 q^{N_1+1} \, (1-q)^{N_2} \, \mathrm{d}q}{\int_0^1 q^{N_1} \, (1-q)^{N_2} \, \mathrm{d}q}$$

Laplace does the same thing (or, rather, we have done the same thing as Laplace),

$$E = \frac{\int x^{p+1}(1-x)^q \, \mathrm{d}x}{\int x^p (1-x)^q \, \mathrm{d}x}$$

All that's left to do is to solve these Dirichlet integrals in the numerator and denominator. The numerator is,

$$\int_0^1 q^{N_1+1} \, (1-q)^{N_2} \, \mathrm{d}q = \frac{\prod_{i=1}^2 \Gamma(N_i + 1)}{\Gamma \left[\sum_{i=1}^2 (N_i + 1) \right]} = \frac{(N_1 + 1)! \, N_2!}{(N + 2)!}$$

In the same way, the denominator is,

$$\int_0^1 q^{N_1} \, (1-q)^{N_2} \, \mathrm{d}q = \frac{\prod_{i=1}^2 \Gamma(N_i + 1)}{\Gamma\left[\,\sum_{i=1}^2 (N_i + 1)\,\right]} = \frac{N_1! \, N_2!}{(N+1)!}$$

The ratio simplifies to,

$$P(A_{N+1} = w \,|\, \mathcal{D}) = \frac{N_1 + 1}{N + 2}$$

Laplace's answer looks like this,

$$E = \frac{p+1}{p+q+2}$$

with $p \equiv N_1$ and $p + q \equiv N$.

Interestingly, Laplace knew the solution for these *Beta* integrals[2], but didn't directly incorporate that into his explanation. Instead, he gives the numerator as,

$$\frac{1 \times 2 \times 3 \times \cdots \times q}{(p+2) \times (p+3) \times \cdots \times (p+q+2)}$$

and the denominator as,

$$\frac{1 \times 2 \times 3 \times \cdots \times q}{(p+1) \times \cdots \times (p+q+1)}$$

with the respective numerators in these expressions canceling, and everything in the respective denominators canceling except for initial $(p+1)$ term and the final $(p+q+2)$ term.

I have to admit that it took me some time to figure out what he was doing here even after knowing what the correct answer must be. Substitute Laplace's notation into the Dirichlet integral solution to see that the numerator is,

$$\frac{\Gamma(p+1+1)\,\Gamma(q+1)}{\Gamma(p+1+1+q+1)} = \frac{(p+1)!\,q!}{(p+q+2)!}$$

and the denominator is,

$$\frac{\Gamma(p+1)\,\Gamma(q+1)}{\Gamma(p+1+q+1)} = \frac{p!\,q!}{(p+q+1)!}$$

to see that Laplace's somewhat cryptic derivation is, in fact, correct.

$$E = \frac{(p+1)!\,q!}{(p+q+2)!} \times \frac{(p+q+1)!}{p!\,q!} = \frac{p+1}{p+q+2}$$

The only reason I bring this up is to make a plea to anyone who has any intent of trying to explain something to his fellow man. Always rely on the simplest and most discursive explanation you are capable of. I know that it will not soothe your vanity, but future generations will thank you.

---

[2]Laplace had been dead for a quarter century before Dirichlet arrived on the scene with his solution.

And so we have arrived at the first version of what become known as Laplace's **Rule of Succession**. According to Laplace, if five white tickets and five black tickets have already been drawn from the urn, the state of knowledge about the next draw being a white ticket is lent a quantitative measure by a probability of,

$$P(A_{N+1} = w \,|\, 5w, 5b) = \frac{N_1 + 1}{N + 2} = \frac{6}{12} = 1/2$$

Note the fact, that, curiously, it doesn't make any difference in your state of knowledge about the next ticket whether *no* tickets have been drawn (our example of what probability to give to the very first flip of the coin), or whether 20,000 white tickets and 20,000 black tickets have already been drawn. But, to use Laplace's language, your state of knowledge about the probability of what *causes* that next ticket to be drawn certainly has changed significantly!

Now, of course, if we use the more general Equation (12.3) with $M_1 = 1$ and $M_2 = 0$, we essentially repeat the same answer given in section 12.5,

$$
\begin{aligned}
P(M_1 = 1, M_2 = 0 \,|\, \mathcal{D}) &= P(M_1 = 1, M_2 = 0 \,|\, N_1, N_2) \\[2ex]
&= C \times \frac{(M_1 + N_1)! \, (M_2 + N_2)!}{M_1! \, M_2!} \\[2ex]
&= C \times (N_1 + 1)! \, N_2! \\[2ex]
C &= \frac{M! \, (N + 1)!}{N_1! \, N_2! \, (M + N + 1)!} \\[2ex]
P(M_1 = 1, M_2 = 0 \,|\, \mathcal{D}) &= \frac{(N + 1)!}{N_1! \, N_2! \, (N + 2)!} \times (N_1 + 1)! \, N_2! \\[2ex]
&= \frac{N_1 + 1}{N + 2}
\end{aligned}
$$

After this, Laplace immediately provided the probability for any number of future drawings of white and black tickets. Unfortunately, he made an error here by forgetting to include the multiplicity factor $W(M)$ in his answer. He corrected this mistake in 1781 where the correct formula is given.

13.7 Connections to the Literature

Feller's classic two volume text [4] was very influential among more rigorously minded statisticians. Those of lesser skill tended to be swayed by his reputation and authority. On a personal note, Feller happened to be one of the very first authors from whom I tried to learn about probability in a more serious manner. Mathematician friends had informed me that if I could ever understand his work, I would then know everything I needed to know about probability.

After failing to keep up with all the mathematical niceties, and not perceiving any kind of connected conceptual theme, I become disenchanted with Feller. I was also heavily influenced by Jaynes's similar negative reaction to Feller's pedagogy.

Jaynes did not particularly care for Feller's disdain for any kind of "Bayesian approach" to inference. Understandably, Feller's perception was heavily moulded by the conventional wisdom as it existed in the 1930s through the 1950s when Feller was in his prime. Feller also attracted Jaynes's ire in his scathing dismissal of Laplace.

Jaynes felt that Feller was a "gamesman" who could not help showing off just how extremely clever he was. Rather than trying to explain how a few fundamental concepts could be widely applied to inferential problems, Feller preferred to astound you with a sophisticated mathematical solution to some problem. More often than not, his solution seemed to be a bolt from the blue, a mathematically creative act, rather than some principled, if prosaic, application of basic probabilistic notions.

Jaynes rebelled against pedagogy as a kind of one–upmanship. He pleaded for some sort of conceptual unity guiding probabilistic inference. In agreement with Jaynes, I have consciously tried to avoid clever solutions in favor of repeating a few basic fundamental notions. So with that ringing in your ears, please bear with me as I now present a very tedious detailed explanation of sample spaces and events whose only virtue is its freedom from mathematical gamesmanship.

Now after many years and untold false leads, it is becoming clearer to me where Feller was right, and where he led the unwary astray. But let me begin on a positive note where Feller's introduction to probability is crystal clear, and easy to follow. My definition of a sample space and events as discussed in this Chapter come directly from his Volume I, Chapter I, pages 7–14. Feller can be mined for a host of useful combinatorial formulas, several of which play a critical role in this Chapter.

So how did Feller choose to explain a sample space? He talked about placing r balls into n cells. He used an example where $r = 3$ balls were placed into $n = 3$ cells. The sample space then consisted of $n^r = 3^3 = 27$ elementary points.[3]

Feller said that the balls were "distinguishable," and labeled them as **a**, **b**, and **c**. He showed a picture of all 27 elementary points in his Table 1 on page 9 with the three balls **a**, **b**, and **c** sitting somewhere in the three cells.

Now this is more easily understood if we make it less abstract. Let's ask whether three of the Halloween party participants would like to help us out here. We will ask Alice to be **a**, Ben to be **b**, and Charlie to be **c**. As separately identifiable individuals, they are playing the role of the "distinguishable balls."

The role of the cells is played by the statements in some state space. Since $n = 3$, say that the state space consists of these three statements about a person's height, "This person is TALL.", "This person is AVERAGE.", and "This person

---

[3]There is an error on page 10, Example (b). For $r = 4$ balls in $n = 3$ cells, the sample space consists of $n^r = 3^4 = 81$ elementary points, not 64 points.

is SHORT." What is the IP's state of knowledge about the height of these three people? Or, what is the probability of any event as defined on this sample space?

An event A might be a *simple event* consisting of just one elementary point, or it could be a *compound event* consisting of an aggregation of many elementary points. The event A could be that, "one cell was multiply occupied," or, in our more transparent example, "at least two people in one of the height categories."

Another event B might be defined as, "first cell not empty," or, "Somebody is TALL." Event C might be that both A and B occur, "Somebody is TALL and at least two people are either TALL, AVERAGE, or SHORT. This could be cell 1 where all three people are TALL, or cells 4 through 15 where at least somebody is TALL, and, if it happens that only one person is TALL, there are two people who are either AVERAGE or SHORT.

The remaining elementary points 16 through 27 describe the complementary event. The cells contain nobody TALL, or if there is one TALL person, the other two people don't fall into the same height category. One can always look at the explicit listing of all the elementary points in this manner, and see exactly which elementary points satisfy any event.

Finally, an "impossible event" might be, "the first cell empty and no cell multiply occupied." For three people, if no one is TALL, there cannot be just one person who is AVERAGE, and one person who is SHORT.

Each elementary point in the sample space is the most refined statement we can make. An example might be, "Alice is AVERAGE, Ben is SHORT, and Charlie is TALL." **a** is in cell 2, **b** is in cell 3, and **c** is in cell 1. This corresponds to cell 26 in Feller's list. But usually we are going to ask about events that are aggregates of these elementary points.

The translation between Feller's notation and mine is fairly direct. The number r of balls corresponds to the future frequency count M. The number of cells, n, corresponds to the number of statements in the state space. Thus, the total number of elementary points in the sample space is $n^r \equiv n^M = 3^3 = 27$. I prefer to emphasize that the goal in constructing the sample space is inferential. An IP wants to assess its state of knowledge about the future frequency counts.

Then Feller introduces the notion that the balls might be "indistinguishable." He reduces the original 27 points in the sample space to only ten points in a new sample space. And, finally, he says that even the cells might be indistinguishable, in which case, the sample space is reduced even further to just three points.

What, till now, has been a perfectly lucid description by Feller becomes a bit murky. Here is a clearer explanation for these notions.

There are only three classes of contingency tables. We could decompose the future frequency count of $M = 3$ into sums such as $3 + 0 + 0$, $2 + 1 + 0$, or $1 + 1 + 1$. All three people could share the same height, two people could have the same

height, and the third a different height, or all three people could have different heights. Feller calls this the case where both balls and cells are "indistinguishable."

Now three people sharing the same height could happen in three ways. This is where we use the counting formula,

$$\frac{n!}{r_z!\, r_s!\, r_d!\, r_t!} = \frac{3!}{2!\, 0!\, 0!\, 1!} = 3$$

These are the contingency tables $\boxed{3\,|\,0\,|\,0}$, $\boxed{0\,|\,3\,|\,0}$, or $\boxed{0\,|\,0\,|\,3}$. Two people with the same height, and the third with a different height can happen in 6 ways,

$$\frac{n!}{r_z!\, r_s!\, r_d!\, r_t!} = \frac{3!}{1!\, 1!\, 1!\, 0!} = 6$$

These are the contingency tables $\boxed{2\,|\,1\,|\,0}$, $\boxed{2\,|\,0\,|\,1}$, $\boxed{1\,|\,2\,|\,0}$, $\boxed{1\,|\,0\,|\,2}$, $\boxed{0\,|\,2\,|\,1}$, and $\boxed{0\,|\,1\,|\,2}$. And, finally, all three people with different heights can happen in just one way,

$$\frac{n!}{r_z!\, r_s!\, r_d!\, r_t!} = \frac{3!}{0!\, 3!\, 0!\, 0!} = 1$$

This is the single contingency table $\boxed{1\,|\,1\,|\,1}$.

The total number of different contingency tables, or frequency counts, is then,

$$3 + 6 + 1 = 10$$

This is where we use the formula,

$$\frac{(M + n - 1)!}{M!\, (n - 1)!} = \frac{(3 + 3 - 1)!}{3!\, 2!} = 10$$

This is what we have called a voluntary act of disregarding information on the part of the IP. The IP does not care which particular person makes up a frequency count. Feller somewhat misleadingly speaks of this as the case of "indistinguishable balls," and shows a list with ten cells in his Table 2.

My criticism here is that, after a completely acceptable explanation of the sample space and events, Feller blunders by stating that the sample space then must be redefined as consisting of only these ten points. I disagree. The sample space remains as it was originally defined. These ten "new sample points" are really just new events defined on the original sample space. The IP doesn't care about which particular people will be involved in this new event.[4]

---

[4]This leads to further confusion for the *(d) Sampling* example on page 12 where the sample space consisting of M=100 people who do or do not smoke ($n = 2$) is said to consist of 101 points. It is true that there are 101 possible frequency counts if the IP doesn't care about the specific individuals involved, but the number of elementary points in the sample space is 2^{100}. Why include the correct description of the sample space for the *(f) Coin tossing* example when according to the "indistinguishable ball" criterion just used for the smoking example, the sample space should consist of just four points, no HEADS through three HEADS?

The same criticism applies to Feller's description of "indistinguishable balls and cells." where he reduces the sample space to just three points. Now the IP doesn't care about which particular people or which particular height, but only cares about the three different breakdowns of the future frequency count $M = 3$ characterized by $3 + 0 + 0$, $2 + 1 + 0$, and $1 + 1 + 1$. Again, this is an aggregation of a certain number of elementary points to define new events.

But when we take account of the "distinguishability" of each person, each one of these contingency tables may happen in more than one way. This is where we use the multiplicity factor. The first type of contingency table with all three people sharing the same trait can only happen in one way,

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3!} = \frac{3!}{3! \, 0! \, 0!} = 1$$

The second type of contingency table where two people have the same height, and the third is different can happen in three ways,

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3!} = \frac{3!}{2! \, 1! \, 0!} = 3$$

The third type of contingency table where all three people have different heights can happen in six different ways,

$$W(M) = \frac{M!}{M_1! \, M_2! \, M_3!} = \frac{3!}{1! \, 1! \, 1!} = 6$$

This kind of decomposition has to add up to the total number of elementary points in the sample space. Each way for the frequency counts when individuals didn't matter, multiplied by the multiplicity factor when they do matter, must equal the total number of elementary points,

$$(3 \times 1) + (6 \times 3) + (1 \times 6) = 27$$

The first term (event) is composed from the three points 1, 2, and 3. The second term (event) is composed from the eighteen points 4 through 21. The third term (event) is composed from the final six points 22 through 27.

There exist three highest level events characterized by both "balls" and "cells" being "indistinguishable." The first event is everybody has the same height, the second event is two people have the same height and the third a different height, and the third event is that everybody has a different height.

We have already discussed the ten lower level events exist where only the "balls are indistinguishable." An example of one of these ten frequency counts (from the second higher level class above) is where two people are SHORT, one person is TALL, and nobody is AVERAGE, $\boxed{1}\,\boxed{0}\,\boxed{2}$.

An example of a lowest level event (issuing again from the same higher level classes), where now both balls and cells are distinguishable, is an elementary point,

say, where Alice and Charlie are SHORT, while Ben is TALL (elementary point 14 consisting of ball **b** in cell 1, no balls in cell 2, and balls **a** and **c** in cell 3.).

The $\mathbf{a}, \mathbf{b}, \mathbf{c}$ that Feller uses is generic. They could refer either to the particular temporal order in which something occurred, the particular color die on which a face appeared, the particular person who was measured for his or her height, and so on. The IP may voluntarily choose to ignore any of this detail when events are defined.

Let me repeat why I have spent so much of your time with this minutiae. It is because, in the end, NOTHING IS MYSTERIOUS. There is no question that you can pose about sample spaces, events, or counts (given, that is, that the state space and the number of future frequency counts have been defined) that isn't perfectly clear and answerable. It is only when the *probability* for these elementary points must be assigned that confusion sets in.

For example, on page 20, Feller says "it is natural to assume that all sample points are equally probable ..." This is true only when one specific model is adopted as we have learned from our careful consideration of the formal rules for probability manipulation. The only way that every point in the sample space has an equal probability is when a very specific model has made a numerical assignment of $1/n$ to each statement in the state space. And, furthermore, the probability for this model must be $P(\mathcal{M}_k) \equiv \delta(q_i - 1/n)$ so that it is, in fact, the *only* model that makes numerical assignments to the statements in the state space.

Thus, the simple event A, elementary point 23, "Alice is TALL and Charlie is AVERAGE and Ben is SHORT." has a probability $P(A) = 1/n^M = 1/27$ *only under the fair model*. According to the formal rules,

$$
\begin{aligned}
P(a, c, b) &= \sum_{k=1}^{\mathcal{M}} P(a \mid c, b, \mathcal{M}_k) \times P(c \mid b, \mathcal{M}_k) \times P(b \mid \mathcal{M}_k) \times P(\mathcal{M}_k) \\[2mm]
&= \int \cdot \int P(a \mid \mathcal{M}_k) \times P(c \mid \mathcal{M}_k) \times P(b \mid \mathcal{M}_k) \times \delta(q_i - 1/n) \, \mathrm{d}q \\[2mm]
&= (1/n) \times (1/n) \times (1/n) \\[2mm]
&= 1/27
\end{aligned}
$$

Likewise, every single point in the sample space has the same probability. The probability for an event is the sum of this probability over the elementary points defining the event. Thus, the event A, "Everybody is the same height." consists of the first three elementary points, and has probability of $P(A) = 1/9$ *under this very definitive model*. The probability for "Everybody has a different height." is six times greater than the probability for "Everybody is TALL." because $P(A) = \frac{W(M)}{n^M}$.

This very definitive model is the operational implementation of saying that the IP is very well–informed indeed. It is so well–informed about the model space that it can summarily dismiss every single model except one. It is very well–informed about Laplace's "probability of causes," in effect saying that there is no doubt whatsoever about "the single cause" for a person's height measurement.

Most importantly, as we have seen in this Chapter, at the other extreme when the IP is "completely uninformed," about these "probability of the causes" things change dramatically. The ten possible frequency counts become equally probable.

The probability for event A, "Everybody is the same height." now becomes $P(A) = 3/10$ as opposed to $P(A) = 1/9$. The event that everybody has the same height is now three times more probable than the event that everybody has a different height. The fact that $\boxed{3\,|\,0\,|\,0}$ can happen in only one way while $\boxed{1\,|\,1\,|\,1}$ can happen in six ways just doesn't matter when the IP has "insufficient reason" to single out one physical cause behind a height measurement for a person. For all the IP knows, the measurements might be taken from a basketball team or attendees at a midgets convention.

We end this discussion of Feller with his most provocative comment of all. After agreeing that, in the end, it might be necessary to assign something other than equal probability to each elementary point, he concurs that a probability of 1/10 could be assigned to each of the ten cases where "indistinguishability" is paramount. But since he didn't approach the issue from the perspective of the formal rules as we did, he could never offer the easy explanation centered on differing states of knowledge about models. Remarkably, he comments that this situation is one that demands the invocation of "Bose–Einstein" statistics.

Jeffreys discussed Laplace's **Rule of Succession**, but didn't like it very much. He felt that the uniform distribution for the prior probability did not do justice to the kind of induction we humans typically engage in. There was insufficient probability attached by Laplace to seeing all or none of one type in a population given that a sufficiently large sample had shown one or the other to be the case.

Jeffreys [14] illustrated his objection with this famous example.

> Thus I may have seen 1 in 1,000 of the "animals with feathers" in England: on Laplace's theory the probability of the proposition, "all animals with feathers have beaks", would be about 1/1000. This does not correspond to my state of belief or anybody else's. [pg. 128]

Suppose Jeffreys had seen, say, 5,000 chickens, ducks, geese, swans, birds, and so forth, and all of these feathered fowl, in his experience, had beaks. Laplace says that the probability of observing that the very next "feathered animal" would have a beak is,

$$P(A_{N+1} \,|\, \mathcal{D}) = \frac{N_1 + 1}{N + 2} = 0.9998$$

But when you ask about not just the very next observation, but whether *all* of a future sample comparable in size to your initial observations, Laplace's rule gives,

$$P(M_1 = 5000, M_2 = 0 \,|\, N_1 = 5000, N_2 = 0) \approx 1/2$$

And then if the population of "animals with feathers" is very large, say, 5,000,000, the probability of seeing every single one with a beak is the 1 in 1000 that Jeffreys

complained about,

$$P(M_1 = 5,000,000, M_2 = 0 \mid N_1 = 5000, N_2 = 0) = \frac{5001}{5,005,001} \approx 1/1000$$

Jeffreys was very unhappy with this result. Even though he hadn't yet seen the whole population of 5,000,000 feathered animals, he felt that the overwhelming evidence from the 5,000 that he *had* seen was a definite indicator that he should adopt a greater degree of belief then Laplace was willing to bestow. He couldn't reconcile the **Rule of Succession**'s answer with the induction that all of us would naturally carry out that all feathered animals do indeed have beaks.

So in a corrective move designed to "fix" Laplace's uniform assessment, Jeffreys placed a greater lump of probability on those two extreme cases where either all, or none of the population possessed some trait. The ramifications from that choice continues to ripple downwards even to the present day. We will have much more to say about the "Jeffreys prior" later.

All I want to comment on at this time is this: Suppose that all competing models have been eliminated such that the probability assigned by the one remaining model to the event in question is some fantastic number like 0.9998. This is what Laplace said was the degree of belief to observe that the next feathered animal had a beak. Now even here, the formal rules of probability, irrespective of anything Laplace had to say, tell you that the probability for obtaining some very large number of future occurrences of this high probability event is still unbelievably low!

For example,

$$P(\text{all have beaks} \mid \text{high probability for one to have beak})$$

$$= \quad W(M) \times 0.9998^{5,000,000}\, 0.0002^0$$

$$\approx \quad 4.59 \times 10^{-435}$$

Once again, we have another reason to appreciate Wolfram's deep insight that predicting the far future is nearly impossible. Even though we have this extremely high probability for the very next feathered animal to possess a beak, it is next to impossible that *all* of the vast number of remaining animals will also have one.

Jeffreys's distaste for Laplace's **Rule of Succession** did not arise, as he claimed, from Laplace's supposedly suspect application of the uniform distribution, but rather from the fact that low probabilities for one event to occur every single time over a large number of trials are exactly what probability theorems tell you must happen! Laplace is not to blame for this.

Jeffreys wanted inference via probability to reproduce an exact deductive result: If an animal has feathers, then it must have a beak. In other words, we have a logical implication. $A \rightarrow B$, A is TRUE, therefore, B must be TRUE. We have already seen that inference via probability can reproduce this deduction when the

proper assignments are made to the joint probability table. A zero must be placed in the cell $A\overline{B}$ where we have the joint statement, "An animal has feathers and it does not have a beak." Then, the probability that an animal has a beak given that it has feathers must be a certainty, $P(B \,|\, A, A \rightarrow B) = 1$. Interestingly, we have allocated some probability to the circumstance of $\overline{A}B$ to cover the duckbill platypus.

No prior observations are needed to confirm this asserted model, and from then on every single animal with feathers must have a beak. This deduction can follow from an inference via probability, but only if the "right" probability assignments are made immediately by some model that itself has a probability of 1. Then, when we ask, as Jeffreys wanted to ask, "If an animal has feathers, does it also have a beak?" inferential reasoning answers, "Yes, it does!"

Jaynes [13] has a most enlightening and provocative discussion of all of this in his whimsically entitled **Monkeys, Kangaroos, and N**. Basically, he comes to the conclusion that it all depends on how the two counting factors in the combinatorial argument are treated.

From a computational point of view, he considers the parameters in the Dirichlet distribution to be of the utmost significance. This is the distribution capturing the state of knowledge about the models that have assigned the supposed "correct" numerical values to the joint probabilities.

This Chapter concludes with a few remarks about attempts to capture a state of knowledge that somehow reflects "complete ignorance." An overwhelming amount of literature exists on this topic which, for the most part and to the detriment of anyone trying to sort things out, produces much heat but sheds little light.

The simple key to understanding much of the debate is to examine the effects of the α_i parameters in the Dirichlet distribution used for the prior probability,

$$P(\mathcal{M}_k) \equiv \mathrm{pdf}\,(q_1, q_2, \cdots, q_n)$$

We have studied in some detail what happens when the α_i are all set equal to 1. All possible models assigning legitimate numerical values to the cells of the joint probability table enter the fray. And no one model is deemed more important than any other. The ratio for any two models always stays at,

$$\frac{P(\mathcal{M}_k)}{P(\mathcal{M}_j)} = 1$$

Returning to the students momentarily, the "fair" model $\mathcal{M}_1 \rightarrow \{1/4, 1/4, 1/4, 1/4\}$ has the same standing as the "trick" model $\mathcal{M}_\star \rightarrow \{0, 0, 0, 1\}$. This "flat," or "uniform" prior has been associated with both Bayes's and Laplace's notion of complete ignorance.

There are two other major landmarks in the literature and controversy surrounding "complete ignorance." The first is Jeffreys's prior where, to keep things simple, we'll revert back to the coin flip scenario and $n = 2$.

Jeffreys felt that the Bayes–Laplace flat prior was not quite right. He suggested as an alternative $\text{pdf}(q, (1 - q)) \propto q^{-1/2}(1 - q)^{-1/2}$. This places more weight on the extreme cases of a trick coin with either both HEADS or both TAILS. Rather than setting $\alpha_1 = \alpha_2 = 1$, if we let $\alpha_1 = \alpha_2 = 1/2$, we have Jeffreys's prior.

Jeffreys seemed to be influenced by physical models[5] where, for example, if a chemical reaction has taken place once, we expect it to take place again with certainty. Everyone, to include both Feller and Jeffreys, seemed to be upset with Laplace's infamous example where, using his ignorance prior, he determined the probability for the Sun to rise tomorrow.

The British scientist J.B.S. Haldane had an even more extreme, and arguably odder, notion of what "complete ignorance" meant. Let $\alpha_1 = \alpha_2$ approach zero from the positive side, so that the relative standing for models would then approach $\text{pdf}(q, (1 - q)) \propto q^{-1}(1 - q)^{-1}$. The probability is concentrated almost exclusively, and split evenly, on either the HEADS trick coin, or the TAILS trick coin. Both the dice rolling and student graduation scenarios also used Haldane's prior.

To complete this very broad brush survey of the α_i parameters in the Dirichlet distribution, let the α_i now march away from $\alpha_i = 1$ in the opposite direction. If the α_i are allowed to proceed in lockstep, getting larger and larger, and finally approaching ∞, then,

$$\text{pdf}(q_1, q_2, \cdots, q_n) \text{ approaches } \delta(q_i - 1/n)$$

Consequently, the probability for any future macrostate F_j,

$$P(F_j) \text{ approaches } \frac{W(M)}{n^M}$$

That is, we return full circle to the counting argument state of knowledge for events in the sample space.

When the parameters for the prior probability are such that the $\alpha_i \to \infty$, events where both counting factors in the combinatorial formula are large receive the highest probability. If the parameters are such that the $\alpha_i = 1$, then the second counting factor, the multiplicity factor $W(M)$, is neutralized and the first counting factor takes over. Finally, as the $\alpha_i \to 0$, both counting factors are neutralized, and the events with the *minimum* multiplicity are emphasized.

Jaynes came to realize that demanding "complete ignorance" in one context may, in fact, serve to promote a corresponding injection of information when viewed from another context. I have included Chapter Fifteen in this introductory Volume that goes into some detail about Jaynes's "kangaroos" and the lessons they teach us about the meaning of "uninformative."

I have made a conscious decision not to point to the voluminous literature that discusses "ignorance priors." It is my opinion that most of this is conceptually confused and misdirected. You could spend an awful lot of your time in these thickets just as lost at the end of your journey as when you first started.

_____

[5] Jeffreys attributes this influence back to Karl Pearson and his *Grammar of Science*.

13.8   Solved Exercises for Chapter Thirteen

Exercise 13.8.1. Provide a detailed discussion of the student graduation scenario.

Solution to Exercise 13.8.1

Suppose we have two statements A and B, where A is the statement we'd like to predict, and B is the statement reflecting whatever is known or controlled. B is included because it is thought to be related to A in some manner, and knowing it should help reduce the uncertainty surrounding A. B is called a "causal factor."

To be more specific, we'll say that $A_N = a_1$ is the statement, "Student N graduates."; while $A_N = a_2$ is the statement, "Student N does not graduate." $B_N = b_1$ is the statement, "Student N scored HIGH on the test."; while $B_N = b_2$ is the statement, "Student N scored LOW on the test."

Suppose further that we'd like to predict future occurrences of A, written as A_{N+1}, A_{N+2}, and so on, up to A_{N+M}. Thus, M indicates how far into the future we'd like to predict; for this example, this means predicting the graduation status for the next M students. If there are no previous data then N becomes zero, and predicting future occurrences involves the very first student, the second student, and so on.

If B is viewed as a causal factor, then we are interested in setting up conditional probability statements for the *next* student, $P(A_{N+1} \mid B_{N+1}, \mathcal{D})$, the probability for A_{N+1} conditioned on knowledge of B_{N+1} as well as knowledge about N pieces of past data. For example, writing $P(A_3 \mid B_3, \{A_1, B_1, A_2, B_2\})$ indicates that we'd like to predict whether the third student graduates given knowledge of the test score. We also happen to know the graduation status and test scores of the first two students. Thus, $M = 1$ and $N = 2$ with $\mathcal{D} \equiv \{A_1, B_1, A_2, B_2\}$.

A and B can each take on only two values where each statement like $(A = a_1)$ can only be TRUE or FALSE. If $(A = a_1)$ is FALSE, then $(A = a_2)$ is TRUE. Previously, we have used A and $\overline{A}$ to indicate this situation. We will want to consider joint statements involving both A and B, so the possibilities increase to four. Therefore, the fundamental state space is set with $n = 4$.

The *joint probability table* reflecting this state space consists of $n = 4$ cells with each cell containing a numerical value assigned to the joint statement. These are the Q_i as determined by some model $\mathcal{M}_k$. Cell 1 indexes the joint statement $(A = a_1 \text{ AND } B = b_1)$, and so on through cell 4 which indexes the joint statement $(A = a_2 \text{ AND } B = b_2)$. The *contingency table* has the same look as the joint probability table, but contains actual frequency counts, not probability assignments.

Bayes's Theorem is applied in the standard manner to solve such questions by setting up the ratio of a joint distribution over a marginal distribution. We derived

a general prediction formula from the formal rules as,

$$P(A_{N+1} \mid B_{N+1}, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \mid B_{N+1}, \mathcal{M}_k) \, P(\mathcal{M}_k \mid \mathcal{D})$$

Both terms in the summation on the right hand side are expanded by Bayes's Theorem.

$$P(A_{N+1} \mid B_{N+1}, \mathcal{M}_k) = \frac{P(A_{N+1}, B_{N+1} \mid \mathcal{M}_k)}{P(B_{N+1} \mid \mathcal{M}_k)}$$

and

$$P(\mathcal{M}_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M}_k) \, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \mid \mathcal{M}_k) \, P(\mathcal{M}_k)}$$

But we are not going to treat things in this complete generality quite yet. We want to start with some simpler situations first. Suppose we begin by saying that there are no past data to rely upon. So N goes away and M begins at 1. Secondly, concentrate on the state of knowledge for the joint distribution of A and B.

Nonetheless, the state of knowledge is still predicated upon the numerical values assigned to the four cells of the joint probability table under some model. The state of knowledge for some specific joint occurrence of graduation and test score for the next M students is then,

$$P(A_1, B_1, A_2, B_2, \cdots, A_M, B_M \mid \mathcal{M}_k)$$

But we are mainly interested in macro–statements, not in these more detailed micro–statements as just written. That is, we want to predict how many students fall into each of the four categories of the contingency table. How many students GRADUATE with a HIGH score? (the first cell in the contingency table.) How many students DO NOT GRADUATE with a LOW score? (the fourth cell in the contingency table.) Thus, we want to express our uncertainty about the future frequency counts M_1 through M_4 where these frequencies must sum to M.

We called these future events F_j, and since we are trying to predict without the benefit of any past data, then,

$$P(F_j) = \sum_{k=1}^{\mathcal{M}} P(F_j \mid \mathcal{M}_k) \, P(\mathcal{M}_k)$$

The F_j are, of course, the future frequency counts $M_1, M_2, \cdots, M_n$.

Exercise 13.8.2. Write out the state space for the student and test score scenario as just presented in the above exercise.

Solution to Exercise 13.8.2

In the generic notation, the state space consists of the $n = 4$ joint statements.

1. "$A = a_1$ AND $B = b_1$."

2. "$A = a_2$ AND $B = b_1$."

3. "$A = a_1$ AND $B = b_2$."

4. "$A = a_2$ AND $B = b_2$."

Previously, we wrote these four statements as AB, $\overline{A}B$, $A\overline{B}$, and $\overline{A}\,\overline{B}$. For this problem, the state space is more clearly defined by the four statements,

1. "Student GRADUATES and obtained a HIGH score."

2. "Student DOES NOT GRADUATE and obtained a HIGH score."

3. "Student GRADUATES and obtained a LOW score."

4. "Student DOES NOT GRADUATE and obtained a LOW score."

Each one of these four statements is either TRUE or FALSE. In other words, each student must be placed into one, and only one, of these four categories.

Suppose the model assigning numerical values were some logic function, say, the `EQUAL` operator as explored in a future exercise, then the state of knowledge about graduation,

$$P(A \,|\, B, A \leftrightarrow B)$$

would be 0 or 1 reflecting certainty of the student's graduation status given the test score. However, other models may generalize away from this logic function in their numerical assignments in order to allow inference to give us a wider range of answers under uncertainty.

Exercise 13.8.3. Sketch the general appearance of a joint probability table for the student graduation scenario. Insert the fair model's, $\mathcal{M}_1$, numerical assignments.

Solution to Exercise 13.8.3

A joint probability table is shown at the top of the next page in Figure 13.3. The table consists of four cells corresponding to $n = 4$ of the state space. The fair model's numerical assignment of $Q_i = 1/4$ is shown inserted into the four cells of the table.

Figure 13.3: *The joint probability table with the numerical assignments and marginal probabilities under model* $\mathcal{M}_1$.

Exercise 13.8.4. What is the appropriate expansion of $P(B \mid \mathcal{M}_1)$ in the denominator of Bayes's Theorem?

Solution to Exercise 13.8.4

Using the **Sum Rule** to expand $P(B = b_1 \mid \mathcal{M}_1)$, we have,

$$P(B = \text{HIGH} \mid \mathcal{M}_1) = P(A = \text{GRADUATES}, B = \text{HIGH} \mid \mathcal{M}_1) +$$

$$P(A = \text{DOES NOT GRADUATE}, B = \text{HIGH} \mid \mathcal{M}_1)$$

This is the same as the marginal probability for $P(B = \text{HIGH} \mid \mathcal{M}_1)$ shown in the joint probability table as $Q_1 + Q_2 = 1/2$.

Exercise 13.8.5. Try to generate a model $\mathcal{M}_2$ that results in numerical assignments such that there *is* an association between A and B. You are restricted to keeping the marginal probabilities the same as before.

Solution to Exercise 13.8.5

Figure 13.4 at the top of the next page presents a different joint probability table arising from numerical assignments under a different model $\mathcal{M}_2$. Substitute these numerical assignments into the first term of the prediction equation,

$$P(A = a_1 \mid B = b_1, \mathcal{M}_2) = \frac{P(A = a_1, B = b_1 \mid \mathcal{M}_2)}{P(A = a_1, B = b_1 \mid \mathcal{M}_2) + P(A = a_2, B = b_1 \mid \mathcal{M}_2)}$$

$$= \frac{Q_1}{Q_1 + Q_2}$$

$$= 2/3$$

Figure 13.4: *The joint probability table with the numerical assignments and marginal probabilities under model $\mathcal{M}_2$.*

The probability that a student GRADUATES, given knowledge that he or she scored HIGH on the test, changes from $1/2$ to $2/3$ under model $\mathcal{M}_2$.

Exercise 13.8.6. Make up another model $\mathcal{M}_3$ that provides numerical assignments such that there is an even stronger association between A and B. Keep the marginal probabilities the same as before.

Solution to Exercise 13.8.6

Substitute the numerical assignments from the information under a changed model $\mathcal{M}_3$ into the first term of the prediction equation to find that,

$$P(A = a_1 \mid B = b_1, \mathcal{M}_3) = \frac{P(A = a_1, B = b_1 \mid \mathcal{M}_3)}{P(A = a_1, B = b_1 \mid \mathcal{M}_3) + P(A = a_2, B = b_1 \mid \mathcal{M}_3)}$$

$$= \frac{Q_1}{Q_1 + Q_2}$$

$$= 5/6$$

Figure 13.5 at the top of the next page shows another joint probability table dictated by a third model $\mathcal{M}_3$. The probability under model $\mathcal{M}_3$ that a student GRADUATES given that he or she scored HIGH on the test is $5/6$, not $1/2$ as under model $\mathcal{M}_1$, nor $2/3$ as under model $\mathcal{M}_2$.

Figure 13.5: *The joint probability table with the numerical assignments and marginal probabilities under model $\mathcal{M}_3$.*

Exercise 13.8.7. Conduct a counting exercise for the graduation scenario sample space.

Solution to Exercise 13.8.7

There is some merit in introducing and discussing at some length, a *combinatorial* approach to the problem as we have just set it up. Combinatorial arguments are the distilled wisdom gained in counting up various ways things might happen. Counting in this way is conceptually distinct from information, probability, and states of knowledge. Nevertheless, it is very instructive to examine the interplay between the often confused concepts of frequency and probability.

We are going to engage in such a counting exercise for the scenario involving the students, their test scores, and graduation. Let interest be focused on the next four students, so set $M = 4$. Let's give these four students names so that we might refer to them as individuals. Their names are Alex (**a**), Beth (**b**), Carl (**c**), and Dawn (**d**).

The state space, as mentioned above, consists of four joint statements, so let $n = 4$ as well. We want to assess our state of knowledge about various frequency counts such as: What is the probability that all four students GRADUATE with a HIGH score? This would be written out more explicitly with the future frequencies specified for each cell of the contingency table,

$$P(F_j) \equiv P(M_1 = 4, M_2 = 0, M_3 = 0, M_4 = 0)$$

Or, we might contemplate even higher level macro–statements such as: What is the probability that all four students have the same characteristics, or traits, or attributes? Thus, this statement encompasses the one just listed, plus three others.

For example, one of these other three statements is the probability that all four students DO NOT GRADUATE with a LOW score and written as,

$$P(F_j) \equiv P(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 4)$$

We shall establish that the fundamental counting space consists of 256 points,

$$n^M = 4^4 = 256$$

These are the elementary points that constitute the *sample space*. Every situation can be described in its lowest level detail by referring to one of these points. For example, one point among the 256 possible points is that Alex, Beth, and Dawn GRADUATE with a HIGH score, and Carl DOES NOT GRADUATE with a LOW score. This is one of the four situations where $M_1 = 3$, $M_2 = 0$, $M_3 = 0$, and $M_4 = 1$. There are 255 more elementary points just like this, and we are going to discuss how to generate and count every one of these points.

Be forewarned that we are now embarking on an extremely tedious and excruciatingly boring enterprise when we attempt to verbally describe this counting process. Feel free to skip this part and return only when the need for a detailed explanation is pressing. We are going to proceed from the bottom up by counting all the possibilities, and then sum them to see if they total to the required 256 points. Then, after we discern the pattern that takes place, we can encapsulate this insight into formulas.

To begin, think of all the ways that the $M = 4$ sum could be broken down into $n = 4$ components. The sum 4 could be partitioned as $4 + 0 + 0 + 0$. This could be achieved with $\boxed{4\,|\,0\,|\,0\,|\,0}$ with $M_1 = 4$ and the other $M_i = 0$. But there are four different ways to retain $\sum_{i=1} M_i = M = 4$. The frequency count of 4 gets distributed in turn to the first cell, second cell, third cell, and, finally, the fourth cell.

The counting formula for building up to the total sum is developed by dividing the process into two stages. The first stage was accomplished above. The second stage involves how many different ways the students considered as individuals could participate in the breakdown. With all four students in one of the cells, there is only one way this can happen. Thus, $4 \times 1 = 4$ elementary points start off the summation process. We only have to account for 252 more elementary points!

The $M = 4$ sum could also be partitioned as $3 + 1 + 0 + 0$. This could be achieved by $\boxed{3\,|\,1\,|\,0\,|\,0}$ with $M_1 = 3, M_2 = 1, M_3 = 0, M_4 = 0$. But there are 6 ways to distribute the two zeroes around the four cells, and then two ways for the 3 and 1 to be interchanged for 12 different ways in all. For example, a $\boxed{0\,|\,3\,|\,0\,|\,1}$ contingency table also has the desired pattern represented as $3 + 1 + 0 + 0$.

And now, given that there are three students in one cell and one student in another cell, there are four different ways for the students to distribute themselves in this fashion. For example, in the first of the 12 configurations, $\boxed{3\,|\,1\,|\,0\,|\,0}$, **abc** could be in cell 1 and **d** in cell 2, or **abd** in cell 1 and **c** in cell 2, and so on. Multiplying the number of ways in the two stages yields $12 \times 4 = 48$ elementary points for a total of $4 + 48 = 52$ toward the grand sum of 256.

How else could the sum be partitioned? $2 + 2 + 0 + 0$ is another possibility. So we have to consider this partition with all of its different configurations which are six in number. There are the six different ways to distribute the two 0s around the four cells, but switching

the 2s doesn't make any difference (or vice versa). However, there are also six ways four students could be distributed two and two. Thus, we have $6 \times 6 = 36$ more elementary points with a running total of $4 + 48 + 36 = 88$.

A pattern is beginning to manifest itself. Think of another way that 4 could be partitioned with four cells available. $2 + 1 + 1 + 0$ is a possibility as exemplified by a frequency count of $\boxed{2\,|\,1\,|\,1\,|\,0}$. There are 12 different ways for this partition to take place while maintaining $M = 4$; six ways to distribute the two 1s around the four cells multiplied by the two ways of exchanging the 2 and the 0.

There are also 12 ways in the second stage for the students as individuals to distribute themselves as two to one cell, one to another, and the final one to a third cell. There is a significant addition of elementary points in this situation, $12 \times 12 = 144$. Here is an example of one of these 144 elementary points, **ab** in cell 1, **c** in cell 2, and **d** in cell 3. The running total is up to $4 + 48 + 36 + 144 = 232$.

There is just one last partition to take account of. It is the partition of 4 into $1+1+1+1$. There can be only one contingency table where all $M_i = 1$, $\boxed{1\,|\,1\,|\,1\,|\,1}$. There is a single student in each of the four cells. There is only one possibility in stage 1 counting. However, during stage 2 counting there are a lot of ways that four students could distribute themselves one to a cell, 24 different ways to be exact. Thus, $1 \times 24 = 24$, and the grand sum of $n^M = 4^4 = 256$ is now decomposed as a set of smaller sums $4 + 48 + 32 + 144 + 24 = 256$.

Based on this discursive explanation for the counting process, a combinatorial formula due to Feller [4] is very useful.[6] The counts making up the decomposition of n^M can be calculated by,

$$\text{Count} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1!\, M_2!\, \cdots M_n!}$$

where r_i is the number counting the *repetition* of the ith frequency count. For example, to compute the different number of ways for the $4 + 0 + 0 + 0$ partition, a frequency count of 0 was repeated 3 times, $r_z = 3$, a frequency count of 1 did not occur, $r_s = 0$, and so on, until a frequency count of 4 was repeated one time, $r_M = 1$. Thus,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} = \frac{4!}{3!\, 0!\, 0!\, 0!\, 1!} = 4$$

The second term in the counting formula is the multiplicity factor,

$$\frac{M!}{M_1!\, M_2!\, \cdots\, M_n!} = \frac{4!}{4!\, 0!\, 0!\, 0!} = 1$$

Thus, we have,

$$\text{Count} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1!\, M_2!\, \cdots M_n!} = 4 \times 1 = 4$$

Here is the same calculation for the $3 + 1 + 0 + 0$ partition. The first counting term is,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} = \frac{4!}{2!\, 1!\, 0!\, 1!\, 0!} = 12$$

---

[6]Feller's Chapter 2, pp. 39–40, discusses this formula in terms of the classical occupancy problem of M balls in n cells. Feller illustrates it with a table showing how seven accidents can be distributed across the seven days of the week. See the Supplemental Exercises to this Volume for my very detailed explanation.

That is, a frequency count of 0 was repeated twice, a frequency count of 1 was repeated once, a frequency count of 2 did not occur, a frequency count of 3 was repeated once, and a frequency count of 4 did not occur. The second counting term is,

$$\frac{M!}{M_1!\, M_2!\, \cdots M_n!} = \frac{4!}{3!\, 1!\, 0!\, 0!} = 4$$

for a total of,

$$\text{Count} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1!\, M_2!\, \cdots M_n!} = 12 \times 4 = 48$$

contributing to the grand total of 256.

We have seen the elements involved in these two stages of counting before. Perhaps most recognizable is the second stage where the different ways the M students could be distributed in M_1, M_2, $\cdots$, M_n ways is the multiplicity factor, $W(M)$. The first stage where the different ways the sum could be decomposed is a bit harder to fathom. But when we determined the number of macro-statements for n and N in the last Chapter (section 12.3), we developed a formula which when applied to this situation calculates to,

$$\frac{(M+n-1)!}{M!\, (n-1)!} = \frac{(4+4-1)!}{4!\, 3!} = 35$$

This is indeed the total number of ways to sum the frequency counts to $M = 4$ in n cells. It is determined in the first counting stage as $4 + 12 + 12 + 6 + 1 = 35$ where the first two terms in this sum were just discussed. An explicit listing of all 35 ways together with their associated multiplicity factors was shown in Figure 13.1 appearing in section 13.4.1. This sketch summarizes and confirms the laborious explanation provided above.

Having accomplished this much, look at the fraction of the elementary points for various macro–statements. This is easily done by placing the counts as just calculated over n^M,

$$\frac{\text{Count}}{n^M}$$

Start with the macro–statement where all students share the same characteristic. That is, all four students GRADUATE with a HIGH score, or all four students DO NOT graduate with a HIGH score, and so on. This situation was described with a $\boxed{4 \mid 0 \mid 0 \mid 0}$ contingency table, or a $\boxed{0 \mid 4 \mid 0 \mid 0}$ contingency table, and so on.

The fraction of these four elementary points over the total number of points in the sample space is 4/256. This is the macro–statement with the smallest fraction. The macro–statement with the largest fraction is 144/256, describing two students with the same characteristic, with the third and fourth having different characteristics. The largest fraction is NOT where all four students have different characteristics which is only 24/256. It is important to emphasize that these kind of efforts are conceptually distinct from probability considerations.

Exercise 13.8.8. Discuss some various models that might be used in the student graduation scenario.

Solution to Exercise 13.8.8

Now let's tie in the combinatorial arguments with the probabilistic development. We have mentioned many times a sort of standard benchmark model that is usually among the first models to be proposed. It is often characterized as the "fair" or "unbiased" model.

Any k^{th} model will assign legitimate numerical values to all four cells of the joint probability table. Suppose that model $\mathcal{M}_1$ is the benchmark model just mentioned, the model analogous to the fair coin or die. All four cells of the joint probability table (indexing all possible occurrences of A and B) will contain the same numerical assignment of $Q_i = 1/4$. This kind of model denies the causal factor any influence over the variable to be predicted. Here, the model says that graduation does not depend on the test score whatsoever.

Bayes's Theorem confirms the first model's benchmark prediction that the test score doesn't make any difference.

$$P(A = a_1 \,|\, B = b_1, \mathcal{M}_1) = \frac{P(A = a_1, B = b_1 \,|\, \mathcal{M}_1)}{P(A = a_1, B = b_1 \,|\, \mathcal{M}_1) + P(A = a_2, B = b_1 \,|\, \mathcal{M}_1)}$$

$$= \frac{Q_1}{Q_1 + Q_2}$$

$$= 1/2$$

The probability of a student graduating under this model is $1/2$, the same as the marginal probability $P(A)$, no matter what test score the student obtained, or even if we don't know the test scores. B's status as a causal factor is under severe strain. $Q_1 = 1/4$ and $Q_2 = 1/4$ represent the numerical values assigned by model $\mathcal{M}_1$ to the joint probabilities required in Bayes's Theorem. They appear in the first two cells of the four cell joint probability table for this model.

The probability of some future macro–statement F_j, conditioned on assuming model $\mathcal{M}_1$ is true, is,

$$P(F_j \,|\, \mathcal{M}_1) = W(M)\, Q_1^{M_1} Q_2^{M_1} \cdots Q_n^{M_n}$$

Let's return to our current example and find the probability for the particular frequency count where all four students GRADUATE with a HIGH score under this initial model.

$$\begin{aligned}
P(F_j \,|\, \mathcal{M}_1) &= \frac{M!}{M_1!\, M_2!\, M_3!\, M_4!}\, Q_1^{M_1} Q_2^{M_1} Q_3^{M_3} Q_4^{M_4} \\[2mm]
&= \frac{4!}{4!\, 0!\, 0!\, 0!}\, (1/4)^4\, (1/4)^0\, (1/4)^0\, (1/4)^0 \\[2mm]
&= 1/256
\end{aligned}$$

The probability for the other three frequency counts where all four students share the same trait is obviously also going to work out to $1/256$. The probability for the higher level statement that all the students share the same trait is therefore $4/256$.

How about another higher level statement where, say, two students share the same trait, while the other two students have different traits? Specifically, consider the situation where two GRADUATE with a HIGH score, one DOES NOT graduate with a HIGH score, and one GRADUATES with a LOW score. The same kind of analysis yields the probability,

$$
\begin{aligned}
P(F_j \,|\, \mathcal{M}_1) &= \frac{M!}{M_1!\, M_2!\, M_3!\, M_4!} \; Q_1^{M_1}\, Q_2^{M_1}\, Q_3^{M_3}\, Q_4^{M_4} \\[2mm]
&= \frac{4!}{2!\, 1!\, 1!\, 0!} \; (1/4)^2\, (1/4)^1\, (1/4)^1\, (1/4)^0 \\[2mm]
&= 12/256
\end{aligned}
$$

And, referring back to the presentation in the counting argument, we learned that there are 12 possibilities for this kind of $2+1+1+0$ partition. The above case is an example of such a pattern. Summing over all 12 statements of this type, we have a probability of 144/256 for the higher level statement posed at the beginning.

It should be clear by now that the generic probability for the frequency count under model $\mathcal{M}_1$,

$$
P(F_j \,|\, \mathcal{M}_1) = W(M) \times Q_1^{M_1} \times Q_2^{M_1} \times \cdots \times Q_n^{M_n}
$$

is always going to be,

$$
P(F_j \,|\, \mathcal{M}_1) = \frac{W(M)}{n^M}
$$

and the probability for the higher level statement is the sum over all the possible ways represented by the first term in the counting formula.

We conclude, therefore, that the fractions derived from the pure counting analysis are exactly the same as the probabilities under the fair model. The multiplicity factor $W(M)$ is seen to be the deciding factor for both the frequencies, and the probability for events.

Carefully note that we are voluntarily discarding information in order to form macro–statements, or events. In the first instance, for what we have called the lower level statements, or events, the information about the actual identity of the students has been discarded. In the second instance, for what we have called the higher level statements, or events, not only has the identity of the students been neglected, but information about the particular graduation and test score characteristics has been discarded as well.

These higher level events are concerned simply with whether all the students share the same characteristic. We don't care whether that means all four students GRADUATE with a HIGH score, or all four students DO NOT GRADUATE with a LOW score. These two statements (and the other two possible statements just like them) are the same as far as the discarded information is concerned. The probability that all four students share the same trait versus the probability that some of them have dissimilar traits is 4/256 versus 252/256 under this model.

This discussion takes place with an analogy to statistical mechanics in mind. Detailed information about, say, the dynamic path of every atom or molecule in a gas is discarded in favor of finding some macro–variable like the pressure of the gas. In principle, that more detailed *micro*–information is available, even though the IP chooses to ignore it for various pragmatic reasons.

What happens when we consider models other than the fair model? Bayes's Theorem just told us that the fair model did not permit the causal factor B to have any influence on the predicted variable A. The test score was useless in predicting graduation. Think of some model $\mathcal{M}_2$ that *does* result in an association between test score and graduation.

Exercise 13.8.5 presented one such model $\mathcal{M}_2$. Now the probability that a student GRADUATES given a HIGH test score is 2/3, a shift away from 1/2 and illustrating A's dependence B. With the changed numerical assignments under the new model, the probabilities for events are no longer the same as the fractions derived by the combinatorial arguments.

For example, under this new model $\mathcal{M}_2$ the probability that all four students graduate with high scores is,

$$
\begin{aligned}
P(F_j \,|\, \mathcal{M}_2) \;\; &= \;\; \frac{M!}{M_1!\; M_2!\; M_3!\; M_4!} \; Q_1^{M_1} \, Q_2^{M_1} \, Q_3^{M_3} \, Q_4^{M_4} \\[2mm]
&= \;\; \frac{4!}{4!\; 0!\; 0!\; 0!} \; (1/3)^4 \, (1/6)^0 \, (1/6)^0 \, (1/3)^0 \\[2mm]
&= \;\; 0.0123
\end{aligned}
$$

as opposed to $1/256 = 0.0039$ under model $\mathcal{M}_1$. Likewise, the probability that all four students share the same trait changes from $4/256 = 0.0156$ to 0.0262 under model $\mathcal{M}_2$.

Now continue this line of thought by thinking of yet another model that results in an *even stronger* association between test score and graduation. Refer back to Exercise 13.8.6 for a third model, $\mathcal{M}_3$, where the probability that a student GRADUATES given a HIGH test score rises to 5/6. The probability that all four students share the same trait changes from 0.0156 under model $\mathcal{M}_1$ to 0.0262 under model $\mathcal{M}_2$, and finally to 0.0603 under model $\mathcal{M}_3$.

Interestingly, and as you might expect, most of that probability of 0.0603 is provided by either all four students graduating with a high score, or all four students not graduating with a low score.

Extrapolate to what seems to be the obvious conclusion. Propose yet another model $\mathcal{M}_4$ that permits the strongest possible relationship between test score and graduation. Bayes's Theorem now tells us that a student is certain to GRADUATE given a HIGH test score or, certain to NOT GRADUATE given a LOW score since the probability of A conditioned on B under this latest model is 1.

The event that all four students share the same characteristic rises to 0.125 with this probability split equally between the two certain conditions. These are $\boxed{4}\,\boxed{0}\,\boxed{0}\,\boxed{0}$ and $\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{4}$ with probability of 1/16 each. The complementary event that the students have dissimilar characteristics drops from a high of $252/256 = 0.9844$ under model $\mathcal{M}_1$ to a low of 0.875 under model $\mathcal{M}_4$. These are $\boxed{3}\,\boxed{0}\,\boxed{0}\,\boxed{1}$, $\boxed{2}\,\boxed{0}\,\boxed{0}\,\boxed{2}$, and $\boxed{1}\,\boxed{0}\,\boxed{0}\,\boxed{3}$ with probabilities of 0.25, 0.375, and 0.25 respectively.

The influence of the multiplicity factor has been increasingly modulated as we progressed through these four models. Also, model $\mathcal{M}_4$ happened to be the EQUAL logic operator, $A \leftrightarrow B$, (the **Xnor** function in *Mathematica*). This logic function assumes the functional value of T under two conditions; when A and B are both T, or when A and B are both F. The correspondence to the current problem is immediate.

Up to this point, we have examined the predictions from four possible models. But this sample of four models from the entire model space is just a drop in the bucket. Just like "trick" coins and dice, we have models for "trick" graduation scenarios,

$$\mathcal{M}_\star = \{Q_1, Q_2, Q_3, Q_4\} = \{1, 0, 0, 0\}$$

describing the situation where every single student GRADUATES with a HIGH score. Or, perhaps, more disturbing to the creators of the test,

$$\mathcal{M}_{\star\star} = \{Q_1, Q_2, Q_3, Q_4\} = \{0, 0, 1, 0\}$$

where every student GRADUATES with a LOW score. Then, there are just sort of arbitrary, but still quite legitimate, models assigning numerical values like,

$$\mathcal{M}_{\star\star\star} = \{Q_1, Q_2, Q_3, Q_4\} = \{0.314674, 0.098976, 0.428555, 0.157795\}$$

And certainly one wouldn't want to slight that equally memorable model,

$$\mathcal{M}_{\star\star\star\star} = \{Q_1, Q_2, Q_3, Q_4\} = \{0.314673, 0.098976, 0.428555, 0.157796\}$$

It is quite a remarkable accomplishment that there even exists an analytical solution when marginalizing over the entire continuous model space. Nonetheless, back in Exercise 12.6.7, we derived the probability for any future frequency count when all possible models were accorded an equal weight,

$$P(F_j) = \int \cdot \int_{\sum q_i = 1} W(M)\, q_1^{M_1} q_2^{M_2} \cdots q_n^{M_n} \, \mathrm{pdf}\,(q_1, q_2, \cdots, q_n)\, \mathrm{d}q_i = \frac{M!\,(n-1)!}{(M+n-1)!}$$

The inverse of this formula tells us how many such contingency tables exist. For the current example, there are,

$$\frac{(M+n-1)!}{M!\,(n-1)!} = \frac{(4+4-1)!}{4!\,3!} = 35$$

possible contingency tables, all of which were shown in Figure 13.1. Each one of these future frequency counts has a probability of 1/35 under complete ignorance.

Thus, 4 0 0 0, (a condensed way of showing the contingency table when all four students GRADUATE with a HIGH score), has the same probability as 0 2 1 1, (a condensed way of showing that no student GRADUATES with a HIGH score, two students DO NOT GRADUATE with a HIGH score, one student GRADUATES with a LOW score, and the final student DOES NOT GRADUATE with a LOW score).

This situation averaging all possible models is in some sense just the opposite of the "fair" model assigning $Q_i = 1/n$. The contingency table 4 0 0 0 had a probability of only 1/256 under the fair model, but 0 2 1 1 had a probability of 12/256. The greater probability of the latter event was due entirely to its greater multiplicity, $W(M) = 12$ as compared to $W(M) = 1$.

The multiplicity factor is seen to be all–important in the fair model, but plays no role whatever when all models are considered equal. Its influence has been completely negated.

The probability for the higher level event where all four students have the same trait is now 4/35. The probability for the complementary event is 31/35 under the Laplacian perspective of an insufficient reason to rely on one cause only. Compare this to 4/256 versus 252/256 when there is a sufficient reason to rely on just one possible cause, the fair model.

Exercise 13.8.9. What would the joint probability table look like if the EQUAL logic function were used as inspiration for a model?

Solution to Exercise 13.8.9

Employ the tried and true tactic of finding the 0s in the joint probability table for this logic function through the function's DNF expansion. We find that, since $(A \wedge B) \vee (\overline{A} \wedge \overline{B})$ is the DNF expansion for $A \leftrightarrow B$, 0s should be placed in cells 2 and 3 of the joint probability table for A and B when this particular logic function is used as the model for making the numerical assignments.

See Figure 13.6 at the top of the next page for the joint probability table that implements the numerical assignments for the EQUAL logic function $A \leftrightarrow B$. Cells 2 and 3 contain 0, and the values in cells 1 and 4 reflect the inherent symmetry in the variable assignment of T and F in the Boolean Algebra. We will henceforth write this as $P(A, B \mid \mathcal{M}_8)$, where $\mathcal{M}_8$ indicates that numerical values have been assigned in consonance with the logic operator $f_8(A, B) \equiv A \leftrightarrow B$.

We have taken every opportunity to demonstrate that probability theory will reproduce any answer obtained from Classical Logic. Using the formal manipulation rule encapsulated in Bayes's Theorem, we must be able to recapitulate any of the certainties of the functional assignments for any of the arguments of this, or any other, logic function.

	A	**$\bar{A}$**	
B	**Q_1** 1/2	**Q_2** 0	**P(B)** 1/2
$\bar{B}$	**Q_3** 0	**Q_4** 1/2	**P($\bar{B}$)** 1/2
	1/2 **P(A)**	1/2 **P($\bar{A}$)**	**1**

Figure 13.6: *Joint probability table with the numerical assignments and marginal probabilities for the model implementing the* `EQUAL` *logic function. The cells contain the joint probabilities* $P(A, B \,|\, \mathcal{M}_8 \equiv A \leftrightarrow B)$.

To that end, show the results from Bayes's Theorem as,

$$P(A = T \,|\, B = T, A \leftrightarrow B) \;\; = \;\; 1$$

$$P(A = T \,|\, B = F, A \leftrightarrow B) \;\; = \;\; 0$$

$$P(A = F \,|\, B = T, A \leftrightarrow B) \;\; = \;\; 0$$

$$P(A = F \,|\, B = F, A \leftrightarrow B) \;\; = \;\; 1$$

In the context of this particular problem, the `EQUAL` logic function is acting like a model asserting that a student must GRADUATE if he or she scores HIGH on the test and, likewise, must NOT GRADUATE if he or she scores LOW. It cannot happen, under this model, that a student GRADUATES after scoring LOW or DOES NOT GRADUATE after scoring HIGH.

Exercise 13.8.10. With the joint probability table just constructed for $A \leftrightarrow B$ as inspiration, generate another model in the model space that is very close to it.

Solution to Exercise 13.8.10

In this case, pick out for special attention two models assigning numerical values for the probabilities of the statements in the state space. The first model is exactly the same as the logic function model, while the second is "very close" to it. Figure 13.7 shows the two joint probability tables side by side.

Under the model of the `EQUAL` logic function as shown in the joint probability table on the left hand side of Figure 13.7, whether a student graduates or not given

	A=G	A=NG	
B=HIGH	Q_1 1/2	Q_2 0	P(B=H) 1/2
B=LOW	Q_3 0	Q_4 1/2	P(B=L) 1/2
	1/2	1/2	1
	P(A=G)	P(A=NG)	

	A=G	A=NG	
B=HIGH	Q_1 .49	Q_2 .01	P(B=H) 1/2
B=LOW	Q_3 .01	Q_4 .49	P(B=L) 1/2
	1/2	1/2	1
	P(A=G)	P(A=NG)	

Figure 13.7: *Two joint probability tables with the numerical assignments and marginal probabilities for 1) a model that matches the* EQUAL *logic function, and 2) another model that is very close to it.*

his or her test score is a certainty. This represents the strongest possible association between the factor to be predicted, the graduation status, and its supposed causal factor, the student's test score.

The model shown in the joint probability table on the right hand side dictates that the probability for graduating given a high test score is not quite as certain as under the first model, but it is nonetheless very close to 1. One can easily imagine a panoply of other possible models as they begin moving away from the model that is this particular logic function.

We have, in fact, already investigated three of these models earlier as models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. For example, model $\mathcal{M}_3$, whose numerical assignments provided a state of knowledge of 5/6 for a student GRADUATING after obtaining a HIGH test score, could be seen as a variation on the EQUAL logic function model in moving downwards from a higher degree of certainty, or, from another perspective, moving upwards from the lower degree of certainty under the fair model.

Exercise 13.8.11. What is the "mirror image" of the EQUAL logic function?

Solution to Exercise 13.8.11

The mirror image of the EQUAL logic function is the XOR logic function. The joint probability table for "exclusive or," $P(A, B \mid \mathcal{M}_9 = A \oplus B)$, has 0s where $A \leftrightarrow B$ has non–zero entries, and *vice versa*. See Figure 13.8 for confirmation.

The DNF for $A \oplus B$ is $(A \wedge \overline{B}) \vee (\overline{A} \wedge B)$. The functional assignment is T whenever the argument values ARE NOT the same, the mirror image of the DNF for $A \leftrightarrow B$ where the functional assignment of T was given when the values of the arguments ARE the same.

Figure 13.8: *Joint probability table with the numerical assignments and marginal probabilities for the model implementing the* XOR *logic function. The cells contain the joint probabilities* $P(A, B \mid \mathcal{M}_9 = A \oplus B)$.

Exercise 13.8.12. What is the joint probability table for the FALSE logic function? What is its mirror image? What are the implications for Bayes's Theorem?

Solution to Exercise 13.8.12

We are asking for an assignment to all four cells of a joint probability table where $\mathcal{M}_1 = A \perp B$. This logic function has F as the functional assignment for all variable assignments. Because T is not assigned anywhere, the DNF has no terms. The DNF is represented as the contradiction, $A \wedge \overline{A}$. Thus, 0s are placed into all four cells of the joint probability table. Already we have a major problem because the cells of the joint probability table must sum to 1. This logic function is the logical contradiction.

Nevertheless, we can still leverage the formal manipulation rules for a consistent and pleasing resolution. It stands to reason that finding a state of knowledge about a student's graduation conditioned on a test score is going to be problematic in this situation. Let's start off with the Bayesian model averaging approach where,

$$
\begin{aligned}
P(A \mid B) &= \sum_{k=1}^{2} P(A \mid B, \mathcal{M}_k) \, P(\mathcal{M}_k \mid B) \\
&= P(A \mid B, \mathcal{M}_1) \, P(\mathcal{M}_1 \mid B) + P(A \mid B, \mathcal{M}_2) \, P(\mathcal{M}_2 \mid B)
\end{aligned}
$$

The "mirror image" function, or dual logic function to $\mathcal{M}_1 \equiv A \perp B$ is the logic function that takes on the functional assignment T for all arguments. It is $f_{16}(A, B) \equiv A \top B$, and this model we take to be model $\mathcal{M}_2$ in the above average.

However, we have a problem with the first term in the above average. We cannot calculate it using Bayes's Theorem because $P(B \,|\, \mathcal{M}_1) = 0$. The impact on Bayes's Theorem is that it is not defined for this situation. Bayes's Theorem does not permit conditioning on a model that represents a logical contradiction.

In addition, we see that such a model represented by all 0s could never allow a student to be categorized in any category. It violates the exhaustion principle that every observation must go into at least one of the state space categories.

Note that it is acceptable to condition on a *statement* that might be FALSE. As a matter of fact, we have done that all along in the example when it happened that $B = b_2$, that is, the causal factor happened to be a LOW score. It is FALSE that $B = b_1$, that the student obtained a HIGH score on the test. This is a legitimate statement to condition on. However, it cannot ever be legitimate to condition on a *model* that is a contradiction.

The mirror image of the `FALSE` logic function, the `TRUE` logic function, has a DNF expansion consisting of four terms. Thus, there are no 0s anywhere in the joint probability table. Each cell might contain the numerical assignment of $Q_i = 1/4$, for example.

The way out is to reconceptualize from two separate four cell joint probability tables to one eight cell joint probability table. Consequently, $\mathcal{M}_1$ and $\mathcal{M}_2$ are simply additional statements whose status is just like A and B. We have three statements, each of which can assume two values for an $n = 8$ dimension state space.

This single joint probability table causes no problems. The first four cells, cells 1 through 4, contain 0s, and the second set of four cells, cells 5 through 8, contain 1/4. The sum of the assignments in all eight cells is 1. With this reorientation, we can return to using the manipulation rules,

$$P(A \,|\, B) \;=\; \frac{P(A, B)}{P(B)}$$

$$P(A, B) \;=\; P(A, B, \mathcal{M}_1) + P(A, B, \mathcal{M}_2)$$

$$\;=\; 0 + 1/4$$

$$P(B) \;=\; P(A, B, \mathcal{M}_1) + P(A, B, \mathcal{M}_2) + P(\overline{A}, B, \mathcal{M}_1) + P(\overline{A}, B, \mathcal{M}_2)$$

$$\;=\; 0 + 1/4 + 0 + 1/4$$

$$P(A \,|\, B) \;=\; 1/2$$

But exactly the same conclusion is reached if the `FALSE` logic function as a model disappeared. If the probability for model $\mathcal{M}_1$ is 0, and there are only two models, then the probability for model $\mathcal{M}_2$ must be 1.

$$P(A \,|\, B) \;\; = \;\; P(A \,|\, B, \mathcal{M}_1) \, P(\mathcal{M}_1 \,|\, B) + P(A \,|\, B, \mathcal{M}_2) \, P(\mathcal{M}_2 \,|\, B)$$

$$P(\mathcal{M}_1 \,|\, B) \;\; = \;\; 0$$

$$P(A \,|\, B) \;\; = \;\; 0 + \left(\frac{1/4}{1/2} \times 1 \right)$$

$$= \;\; 1/2$$

I cannot resist the temptation to characterize the tone of this argument with much the same words as Laplace characterized the Reverend Bayes's exposition, *"une idée très ingénieuse, mais un peu embarrassée."* You might want to revisit Exercise 8.6.6.

Exercise 13.8.13. Think of another way that takes advantage of the formal manipulation rules to make the point of the previous exercise.

Solution to Exercise 13.8.13

The 16 logic functions stand as 16 readily identifiable peaks in the overall model landscape. We have examined two logic functions in particular in these exercises, the EQUAL operator and the XOR operator. They serve as two unique models among the infinite number of models assigning legitimate numerical values to the four cells of a joint probability table.

One could actually classify all the models into these 16 categories with each logic function being the prime exemplar of that category. All the other models in each of these categories would shade gradually away from the mountain peak of the particular logic function that defines the category. Eventually, we would see that a model started bumping up against another mountain peak of some neighboring logic function. We would switch over to thinking of that neighboring logic function as the inspiration for the model.

In any case, one can leverage the formal manipulation rules to gain a different perspective on why inference does not condone conditioning on a contradiction. Start off with some sort of vague intuitive notion that the symmetry of the 16 logic functions might require that $P(A = T, B = T) = 1/4$ when averaged over all models. After all, $A = T$ and $B = T$ always remain as one of the four possible argument settings for two variable logic functions.

In addition, such a notion of symmetry would have the implication that,

$$P(A = T) = P(B = T) = 1/2$$

and finally that,

$$P(A = T \,|\, B = T) = \frac{P(A = T, B = T)}{P(B = T)} = 1/2$$

These stipulations seem reasonable for the abstract notion of a logic function.

Now the formal rules allow us, in the standard way we have been using all along, to say this about $P(A, B)$,

$$P(A, B) \;=\; \sum_{k=1}^{\mathcal{M}} P(A, B, \mathcal{M}_k)$$

$$P(A, B, \mathcal{M}_k) \;=\; P(A, B \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

$$P(A, B) \;=\; \sum_{k=1}^{\mathcal{M}} P(A, B \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

And thus we arrive at the probability for the variable settings as stipulated above,

$$\sum_{k=1}^{16} P(A = T, B = T \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k) = 1/4$$

Expand this sum so that we can explicitly see that,

$$P(A, B) = P(A, B \,|\, \mathcal{M}_1)\, P(\mathcal{M}_1) + P(A, B \,|\, \mathcal{M}_2)\, P(\mathcal{M}_2) + \cdots + P(A, B \,|\, \mathcal{M}_{16})\, P(\mathcal{M}_{16})$$

Invoke symmetry at a different level, namely, Laplace's notion of a symmetry involving the probability of causes. If none of the logic functions has any special status *vis-à-vis* the others, then the probability for any of the models is the same.

$$P(A, B) = \frac{1}{16}\, [\, P(A, B \,|\, \mathcal{M}_1) + P(A, B \,|\, \mathcal{M}_2) + \cdots + P(A, B \,|\, \mathcal{M}_{16})\,]$$

The models have been listed in this order to correspond with the logic functions as given in Chapter 2. $\mathcal{M}_1$ is the contradiction, $\mathcal{M}_5$ is the AND operator, $\mathcal{M}_8$ is the EQUAL operator, and so on. If we now go ahead and insert the numerical assignments for the first cell in the joint probability table where $A = T$ and $B = T$ under each of these 16 models, we have,

$$P(A = T, B = T) = \frac{1}{16}\, [\, 0 + 0 + 0 + 0 + 1 + 0 + 0 + 1/2 +$$

$$0 + 1/2 + 1/2 + 0 + 1/3 + 1/3 + 1/3 + 1/4\,]$$

$$= \frac{1}{16} \left[\, \frac{15}{4}\, \right]$$

For example, the eighth entry in the above sum,

$$P(A = T, B = T \,|\, \mathcal{M}_8) \equiv P(A = T, B = T \,|\, A \leftrightarrow B) = 1/2$$

and the other values can be figured out in the same way. But $P(A = T, B = T)$ clearly does not equal $1/4$.

However, we observe that if we delete the first model from consideration, the sum doesn't change. But now we are averaging over only 15 models, and our requirement is met. Of course, that first model is $A \wedge \overline{A}$, the FALSE logic operator which we have called the contradictory model. So if we set $P(\mathcal{M}_1) = 0$, we are preventing the contradiction from ever appearing as a model, as well as satisfying our symmetry requirements.[7]

Exercise 13.8.14. How would you rearrange the model probabilities for the 16 logic functions in the previous exercise while trying to maintain a close analogy to the Haldane prior?

Solution to Exercise 13.8.14

Recall that we started out by considering all 16 models. But after eliminating the contradictory model, we were left with 15 models. Then we found that,

$$P(A = T, B = T) = 1/4$$

as we had surmised intuitively from the symmetry inherent for the arguments to the logic functions.

What if we now eliminate, as Haldane would like us to do, eleven more models, and settle for only four models? These four models $\mathcal{M}_2$, $\mathcal{M}_3$, $\mathcal{M}_4$, and $\mathcal{M}_5$ represent the four logic operators $\downarrow$, $\star$, $\diamond$, and $\wedge$. Looking at their respective expansions via the DNF, we see that each of these four functions has just one term where the functional assignment is T.

Thus, we would place a 1 as the numerical assignment in the appropriate cell of the joint probability table, and zeroes in the rest of the cells. For example, model $\mathcal{M}_2$, the NOR operator, has a DNF expansion of $\overline{A} \wedge \overline{B}$. The numerical assignment of $Q_4 = 1$ would be placed into cell 4 with the other three cells given an assignment of 0.

If these four logic functions were to split the total model probability equally through the following assignment,

$$P(\mathcal{M}_2) = P(\mathcal{M}_3) = P(\mathcal{M}_4) = P(\mathcal{M}_5) = 1/4$$

then the other twelve models representing the remaining logic functions would not be counted at all because their respective $P(\mathcal{M}_k)$ have been set to 0.

Suppose you wanted to predict some arbitrarily large number of future frequency counts for the joint statement A and B when each statement can happen in only two ways. Hearkening back to the arbitrarily large number for Jeffreys's feathered animals with beaks, pick a total of $M = 5,000,000$ future frequency counts. What is the probability that A and B are both TRUE in $5,000,000$ subsequent observations?

[7]Hofstadter, on the subject of contradictions, says that, "they infect the whole system like an instantaneous global cancer." [10, pg. 196].

In other words, given that AB appears in cell 1, $\overline{A}B$ in cell 2, $A\overline{B}$ in cell 3, and $\overline{A}\,\overline{B}$ in cell 4 of the *contingency table*, what is,

$$P(M_1 = 5,000,000, M_2 = 0, M_3 = 0, M_4 = 0)?$$

Under model $\mathcal{M}_5$, the $\wedge$ operator, the numerical assignment in cell 1 is $Q_1 = 1$, and the assignment to the other three cells is equal to 0. Shortening the probability for the future frequency count to $P(F_j)$, we have,

$$
\begin{aligned}
P(F_j) &= \sum_{k=2}^{5} P(F_j \mid \mathcal{M}_k)\, P(\mathcal{M}_k) \\[2mm]
&= \sum_{k=2}^{5} W(M)\, Q_1^{5,000,000} \times \cdots \times Q_4^0 \times P(\mathcal{M}_k) \\[2mm]
&= 1 \times (0 \times 1/4) + (0 \times 1/4) + (0 \times 1/4) + (1 \times 1/4) \\[2mm]
&= 1/4
\end{aligned}
$$

In other words, $A = T$ and $B = T$ can not happen under models 2, 3, and 4, but must happen under model 5. But the IP was "totally ignorant" about which model was the correct model, and was forced to average the predictions under each of the four models. Thus, the probability is 1/4 for 5,000,000 future occurrences where $A = T$ and $B = T$, and no occurrences of any of the other three possibilities.

Likewise, as we have seen for the Haldane conceptualization of total ignorance, the probability must also be 1/4 for 5,000,000 future occurrences of $A\overline{B}$, and so forth.

What is the probability for the event that all future occurrences are the same? That is, all future frequency counts must all be one of four types: (1) $A = T$ and $B = T$, (2) $A = F, B = T$, (3) $A = T, B = F$, or, (4) $A = F$ and $B = F$. This prediction has a probability of 1. The consequence is that all future occurrences must be of the same kind.

Haldane actually went all the way to the extreme conclusion that Jeffreys wanted when he complained about Laplace's uniform prior over models. It is sufficient to have seen something happen once to predict that it then must happen in every future occurrence. Prior to that first observation, the IP may be said to be "totally ignorant" about which particular causal linkage will take place.

These are the ramifications from Haldane's notion that an IP is completely uninformed. The "maximally uninformed" part about the models paradoxically co–exists with the strongest possible causal linkage between the statements. It is captured operationally by letting the α_i parameters in the Dirichlet distribution for the models approach 0.

Parenthetically, no matter how reasonable or bizarre you might find Laplace's, Jeffreys's, or Haldane's concept of "total ignorance," all are merely special cases within the general template as provided through the formal manipulation rules of probability theory.

Exercise 13.8.15. Use the counting formulas to determine the number of elementary points in an event where two students have the same traits, and the remaining two students each have different traits from the first two and from each other.

Solution to Exercise 13.8.15

The number of elementary points comprising this portion of the $n^M = 256$ total elementary points in the sample space can be calculated by the formula shown in this Chapter as,

$$\textbf{Elementary points} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1! \, M_2! \, \cdots M_n!}$$

For this event, calculate the first term, the different number of ways for the generic 2+1+1+0 split for a sum of 4, where the frequency count for 0 was repeated once, the frequency count for 1 was repeated twice, the frequency count for the 2 was repeated once, and the frequency counts for 3 and 4 did not occur. So,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} = \frac{4!}{1! \, 2! \, 1! \, 0! \, 0!} = 12$$

The second term, the multiplicity factor $W(M)$, is the number of ways to distribute two students to a cell, the remaining two students each to a different cell so that one of the cells remains empty.

$$\frac{M!}{M_1! \, M_2! \, \cdots M_n!} = \frac{4!}{2! \, 1! \, 1! \, 0!} = 12$$

Thus, we have,

$$\textbf{Elementary points} = 12 \times 12 = 144$$

There are 144 elementary points that constitute this event, and it happens to be the most numerous count of all such events.

Exercise 13.8.16. Provide an example of one of these 144 elementary points as discussed in the previous exercise.

Solution to Exercise 13.8.16

An example of one of these 144 elementary points is: Carl DOES NOT GRADUATE with a HIGH score, Beth GRADUATES with a LOW score, and Alex and Dawn both DO NOT GRADUATE with LOW scores. The frequency count of $\boxed{0}\,\boxed{1}\,\boxed{1}\,\boxed{2}$

being one of the 12 ways to obtain the event in the first term, and $\boxed{\star}$ $\boxed{\text{Carl}}$ $\boxed{\text{Beth}}$ $\boxed{\text{Alex, Dawn}}$ one of the 12 possible ways for the four students to distribute themselves for this particular contingency table.

Exercise 13.8.17. What is an IP's state of knowledge concerning the event in Exercise 13.8.15 if it assumes the fair model?

Solution to Exercise 13.8.17

In order to obtain a quantitative measure of its state of knowledge, the IP must calculate the probability for some future event. Therefore, the IP must make an *inference* using the formal rules of probability theory. The IP would like to predict the characteristics of four future students, so M is set to 4 with $\sum_{i=1}^{4} M_i = 4$.

The question posed is to find the probability that two students have the same characteristic, while the remaining two each possess different characteristics, not only from each other but from the first two. This situation is going to be a sum over lower level events. One of these 12 possibilities, as illustrated in the last exercise was, $M_1 = 0$, $M_2 = 1$, $M_3 = 1$, and $M_4 = 2$.

This certainly satisfies the condition with two students NOT GRADUATING with LOW scores, while of the remaining two, one DOES NOT GRADUATE with a HIGH score and the other GRADUATES with a LOW score. The fair model $\mathcal{M}_1$ dictates that all four cells be assigned $Q_i = 1/4$.

$$P(M_1 = 0, M_2 = 1, M_3 = 1, M_4 = 2 \,|\, \mathcal{M}_1) \;=\; W(M) \, Q_1^{M_1} Q_2^{M_2} Q_3^{M_3} Q_4^{M_4}$$

$$W(M) \;=\; \frac{4!}{0! \; 1! \; 1! \; 2!}$$

$$=\; 12$$

$$P(M_1 = 0, M_2 = 1, M_3 = 1, M_4 = 2 \,|\, \mathcal{M}_1) \;=\; 12 \times (1/4)^0 \, (1/4)^1 \, (1/4)^1 \, (1/4)^2$$

$$=\; 12/256$$

The probability would be the same for all other eleven lower level events where we have a $2 + 1 + 1 + 0$ partition, so the probability for the higher level event is $144/256$.

Calculate the count ratio for this future event over the total number of $n^M = 256$ elementary points as $144/256$. It is important to take notice of the fact that this equality between the probability and counting is true only for the fair model, that is, the model assigning $Q_i = 1/n$.

As this exercise illustrates, $\prod_{i=1}^{n} Q_i^{M_i} = n^{-M}$. After the multiplicity factor $W(M)$ multiplies this term, the probability for one of the possible frequency counts satisfying the stated conditions is going to be $W(M)/n^M$. Here there are 12 such contingency tables.

Exercise 13.8.18. Using the same digressive style of this Chapter, talk about rolling three dice.

Solution to Exercise 13.8.18

An IP would like to "predict the future." By now, it has come to realize that "predicting the future" must be an inferential exercise that relies upon the formal manipulation rules of probability. In order to begin, it must first clearly define the state space, and then the sample space. The happy consequence from making these definitions is that no mystery surrounds the meaning of an "event."

The state space for rolling dice has already been discussed, so we set $n = 6$. The IP is interested in the possible future occurrences for three trials. This could mean that one and the same die is rolled the first time, then again the second time, and finally the third time. Or, it could mean that a green die, a red die, and a yellow die are all rolled simultaneously.

In any case, $M = 3$. The sample space will then consist of $n^M = 6^3 = 216$ elementary points. Any event, from a simple event to a compound event, must consist of one or more of these elementary points. In a nod to Feller, we use language that says that three distinguishable balls, labeled **a**, **b**, and **c**, are placed into six distinguishable cells.

An elementary point is the most refined statement we can make. Here, an elementary point might be the statement, "The green die shows a SIX, the red die shows a TWO, and the yellow die shows a FOUR." A higher level statement, a compound event, might be, "All three dice show the same face."

At the highest level, events reflect the partition of 3 into its constituent sums. There are only three such events, $3 + 0 + 0 + 0 + 0 + 0$, $2 + 1 + 0 + 0 + 0 + 0$, and $1 + 1 + 1 + 0 + 0 + 0$. The first event is the statement, "All the dice show the same face." The second event is, "Two of the dice have the same face and the third is different." The third event is, "The dice show different faces."

A contingency table, or frequency count, from each one of these three events might be $\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{3}\,\boxed{0}\,\boxed{0}$, "All three dice show a FOUR.", $\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{2}\,\boxed{1}$, "Two dice show a FIVE and one shows a SIX.", and $\boxed{0}\,\boxed{1}\,\boxed{0}\,\boxed{1}\,\boxed{0}\,\boxed{1}$, "One die shows a TWO and one die shows a FOUR and one die shows a SIX."

There are a total of 56 such contingency tables,

$$\textbf{Number of contingency tables} = \frac{(M + n - 1)!}{M!\,(n-1)!} = \frac{8!}{3!\,5!} = 56$$

three of which have just been shown above.

Of this total, how many belong to the three high level events? For the first event there are 6 contingency tables,

$$\frac{n!}{r_z!\,r_s!\,r_d!\,r_t!} = \frac{6!}{5!\,0!\,0!\,1!} = 6$$

For the second event there are 30 contingency tables,

$$\frac{n!}{r_z!\, r_s!\, r_d!\, r_t!} = \frac{6!}{4!\, 1!\, 1!\, 0!} = 30$$

and for the third event there are 20 contingency tables,

$$\frac{n!}{r_z!\, r_s!\, r_d!\, r_t!} = \frac{6!}{3!\, 3!\, 0!\, 0!} = 20$$

This breakdown must sum to the total number of contingency tables,

$$6 + 30 + 20 = 56$$

If we take account of the fact that the dice are distinguishable through their color, then each one of the contingency tables in the above three classes can happen in a number of ways as reflected in the multiplicity factor. There is only one way that a contingency table like $\boxed{0\ 0\ 0\ 3\ 0\ 0}$ can be formed. The green die, the red die, and the yellow die all show a FOUR.

$$W(M) = \frac{M!}{M_1!\, M_2! \cdots M_6!} = \frac{3!}{0!\, 0!\, 0!\, 3!\, 0!\, 0!} = 1$$

There are three ways that a contingency table like $\boxed{0\ 0\ 0\ 0\ 2\ 1}$ can be formed. The green die and the red die show FIVE and the yellow die a SIX, the green die and the yellow die show FIVE and the red die a SIX, the red die and the yellow die show FIVE and the green die a SIX.

$$W(M) = \frac{M!}{M_1!\, M_2! \cdots M_6!} = \frac{3!}{0!\, 0!\, 0!\, 0!\, 2!\, 1!} = 3$$

There are six ways that a contingency table like $\boxed{0\ 1\ 0\ 1\ 0\ 1}$ can be formed.

$$W(M) = \frac{M!}{M_1!\, M_2! \cdots M_6!} = \frac{3!}{0!\, 1!\, 0!\, 1!\, 0!\, 1!} = 6$$

Multiplying the number of contingency tables in each of the three classes by its multiplicity factor must reproduce the total number of elementary points in the sample space,

$$(6 \times 1) + (30 \times 3) + (20 \times 6) = 216$$

Having established what is clear, and about which there can be no controversy, the IP must now depart from mere counting to an epistemological realm. How does the IP deal with the uncertainty surrounding Laplace's "probability of the causes."

In one common case, the IP can postulate that the dice are, in fact, fair. Fair dice are a justifiable "cause" for a model assigning $Q_i = 1/6$. On the other hand, if the IP cannot convince itself that it knows this much about the dice, then it might adopt a stance of "total ignorance" about the dice. Total ignorance about

the causes for the dice rolls leads to Laplace's uniform prior over models. A uniform prior over model space means that every possible probability will be assigned to the dice.

We have set it up such that the IP is under two polar opposite frames of mind about the causal environment for the dice rolls. When the IP is certain that the dice are fair, the probability for seeing all three dice show the same face is 6/216.

The IP, however, may be maximally uncertain about these dice. The dice may be trick dice, they may have been manufactured with some unknown sorts of physical asymmetries, the person rolling the die may have some "special skill," and so on, and so forth. These constitute "sufficient reasons" to think that there could be several "causes" for the appearance of the spots when the dice are rolled. The IP, in its state of total ignorance, must treat all of these potential causes without favoring one over the other.

The IP follows the advice of Laplace, and assigns equal probability to every single conceivable model making numerical assignments to the six statements in the state space. The probability for seeing all dice show the same face is then much higher at 6/56 as opposed to 6/216.

Exercise 13.8.19. What is the relationship between predicting the next trial given a causal factor, and the general prediction formula?

Solution to Exercise 13.8.19

This is the distinction captured by the two expressions,

$$P(A_{N+1} \mid B_{N+1}, \mathcal{D}) \text{ and } P(M_1, M_2, \cdots, M_n \mid \mathcal{D})$$

The second expression, the posterior predictive probability for future frequency counts, appears in the Bayes's Theorem solution for the first expression,

$$P(A_{N+1} \mid B_{N+1}, \mathcal{D}) = \frac{P(A_{N+1}, B_{N+1} \mid \mathcal{D})}{P(B_{N+1} \mid \mathcal{D})}$$

The choice of $N + 1$ as a subscript is explained because the IP is interested in the very next trial after the occurrence of the N data. Thus, $M = 1$ and $N + M$ becomes $N + 1$.

The joint statement (A_{N+1}, B_{N+1}) is indexed by cell 1 of the joint probability table, while the joint statement $(\overline{A}_{N+1}, B_{N+1})$ is indexed by cell 2. Since the denominator of Bayes's Theorem is,

$$P(B_{N+1} \mid \mathcal{D}) = P(A_{N+1}, B_{N+1} \mid \mathcal{D}) + P(\overline{A}_{N+1}, B_{N+1} \mid \mathcal{D})$$

the future frequency counts,

$$M_1 = 1, M_2 = 0, M_3 = 0, M_4 = 0$$

and,
$$M_1 = 0, M_2 = 1, M_3 = 0, M_4 = 0$$

are inserted into the right hand side of Bayes's Theorem.

Thus, we have the relationship between the two expressions,

$$P(A_{N+1} \,|\, B_{N+1}, \mathcal{D}) =$$

$$\frac{P(M_1 = 1, M_2 = 0, M_3 = 0, M_4 = 0 \,|\, \mathcal{D})}{P(M_1 = 1, M_2 = 0, M_3 = 0, M_4 = 0 \,|\, \mathcal{D}) + P(M_1 = 0, M_2 = 1, M_3 = 0, M_4 = 0 \,|\, \mathcal{D})}$$

Chapter 14

Predicting College Success When Data Are Available

14.1   Introduction

We would like to retain the same general scenario as in the previous Chapter, but make it more realistic. By more realistic it is meant that more causal factors are proposed, and models are allowed to update because data are available.

Thus, an Information Processor might like to predict graduation outcomes for students who have taken *three* tests instead of just one. The IP would also like to take advantage of any past data available on students who have taken these same three tests, and whose graduation status has already been determined.

Unfortunately, trouble arises rather quickly when attempting to generalize from the introductory examples. Probabilistically speaking, we get into trouble because the size of the joint probability tables required by Bayes's Theorem begin to grow exponentially large rather sooner than we would prefer. Upcoming combinatorial exercises illustrate quite nicely what happens here.

This is another verification of Wolfram's observations about ontological systems. Just as it is impossible to predict the micro–behavior of, say, a cellular automaton governed by Rule 110 in its evolution over time, so it is also impossible to predict the fine details of our future, no matter how simple some true underlying physical model of the universe might be.

It then becomes apparent that inferencing must somehow take place at a higher level than trying to predict micro–behavior. Rather prosaically, all this means is, that at the lowest levels, the probabilities are so small as to be meaningless. Instead, we might want to consider elementary points in a sample space after they have been aggregated. The hope is then that these aggregated events accrue some pragmatically meaningful probability.

In the past, this was taken to mean that the single elementary points, while different in their detailed micro–structure, all had the same outward appearance for all we could tell. In other words, we are talking about classical thermodynamics where the detailed motions of myriad atoms and molecules combined in the end to yield the same temperature.

The moral seems to be that, yes, the micro–behavior of a cellular automaton can not be predicted into the far future. All is not lost, however. Suitably defined macro–structures that evolve in the CA might be predictable. These would be perceived, not through deduction as Wolfram would have it, but rather through inferencing.

It might be that the ultimate reward for using inference as opposed to deduction is that prediction about *some aspects* of the future behavior of cellular automata becomes possible. CA are the simplest, most stripped down version of ontological models of how the world really works. Wolfram actually uses variants of CA to model fundamental space and time. We would be most happy indeed if, by forsaking the inconsequential minutiae of life, we could predict suitable macro–structures in the real world *before* their actual appearance in the evolving ontological system.

14.2   Predicting Graduation under More Interesting Conditions

Extrapolate a bit from the conditions laid out in the last Chapter. We are still concerned with predicting the graduation fate of students. It is thought that the results of some test scores will help in that prediction. But now the students have taken three tests, with the test results still broken down rather crudely into HIGH or LOW scores. We have increased the number of causal factors from one to three.

In other words, new statements B, C, and D stand for the three tests, and statement A once again stands for graduation. Each of the four statements A, B, C, and D can take on just two values; namely, GRADUATES or DOES NOT GRADUATE for statement A, and HIGH or LOW for statements B, C, and D. Do not confuse the *statement D* about the third test with the *statement $\mathcal{D}$* concerning the past data.

The state space increases from $n = 4$ to $n = 16$. There are sixteen possible joint statements of the form $(A = a_i, B = b_j, C = c_k, D = d_l)$ in the state space. One example might be,

$$
\begin{aligned}
(A = a_2) &\equiv \text{ DOES NOT GRADUATE} \\[6pt]
(B = b_2) &\equiv \text{ LOW ON FIRST TEST} \\[6pt]
(C = c_1) &\equiv \text{ HIGH ON SECOND TEST} \\[6pt]
(D = d_2) &\equiv \text{ LOW ON THIRD TEST}
\end{aligned}
$$

In principle, each one of these 16 statements in the state space can eventually be judged either TRUE or FALSE when the appropriate measurement has been made. There is an associated joint probability table consisting of sixteen cells, each of which contains an IP's degree of belief about those joint statements. Each cell is assigned a numerical value Q_1 through Q_{16} reflecting the information inserted by some k^{th} model.

The probability is a measure of the information processor's state of knowledge as to whether a joint statement is TRUE. We have seen many examples where a logic function will rule out a particular joint statement by assigning it a value of 0. It is impossible for that joint statement to be TRUE under the given logic function.

But for the purposes of making general inferences, we have seen that the logic functions are too restricted as models. The model space must be expanded to allow for any legitimate numerical assignment of values to the 16, or, in general, to all n cells of the joint probability table.

14.2.1   Recapitulation under the no–data condition

To begin, we will repeat from the last Chapter how predictions get made in the absence of data. The future number of students whose graduation is of interest stays put for now at $M = 4$. However, the state space has increased from $n = 4$ to $n = 16$, so there are far more elementary points in this new scenario. Namely, there are now $n^M = 16^4 = 65,536$ elementary points in the sample space compared to the 256 points examined in the last Chapter.

Nonetheless, it is still not overly onerous to count up the major aggregations. Under the "fair" model where each cell of the joint probability table is assigned a numerical value of $1/16$, the probability of seeing all four students with the same characteristics is very small. The IP's degree of belief is much greater that the four students will possess dissimilar characteristics under this model.

For example, under such a model, it is highly unlikely that we are going to see all the students graduate with high test scores on all three tests. Rather, we are far more likely to observe that the four students have different patterns of graduation and test scores.

Thus, it is likely that we will observe something on the order of: one student graduates with all high scores, one student does not graduate with all low scores, another student graduates with two high scores and a low score, and the final student does not graduate with two low scores and a high score.

When attention is shifted away from a dogmatic single model like the fair model in order to allow for models assigning all possible legitimate numerical values to the cells of the joint probability table, the probabilities for future counts will change.

In doing so, we discover that, rather surprisingly, observing all four students with the same characteristic is just as likely as all four of them having dissimilar characteristics. Any possible future frequency count of four students is equally likely. This situation crops up when the IP is completely uninformed.

And, if we start to allow models of the Jeffreys and Haldane variety, then some models that strongly link characteristics become more prominent. Under these models we would not be surprised to see every single student graduating, and every single student achieving a high score on every test ever administered.[1] The linkage between these characteristics is just too strong for any kind of variability to manifest itself.

As might be suspected, this kind of idle speculation about the relative status of models becomes moot after observations start to be made. For example, equality of models, while completely justified prior to any measurements, is no longer true once that very first data point is collected. The observations then become the driving force behind the information processor's updating of its state of knowledge about models. Equality of causes is then pared away, and the focus shifts to far fewer "causal" models for why students graduate.

14.2.2   Past data are available

Recall the distinction between N, the total number of observations already made, and M, the total number of observations yet to be made. Suppose now that we have in our possession data from $N = 32$ previous students. We would like to predict the graduation status of the *next* M students. A contingency table containing plausible, but nonetheless made–up, data is presented in Figure 14.1.

The $N = 32$ students representing the observed data have been correctly placed into the 16 categories of the contingency table. For example, the four students who happened to GRADUATE with a LOW score on test B, a HIGH score on test C, and a LOW score on test D are shown in the 7th cell. Of the 32 students, 24 students graduated and 8 students did not graduate. The other marginal sums shown for the tests support the idea that HIGH and LOW are meaningful. That is, there are 16 students in the HIGH and LOW category for each of the three tests.

Suppose further that we are still interested in predicting the characteristics for four future students, thus keeping $M = 4$. We must take advantage of the entire model space to begin. Fortunately, we have worked out the prediction equation for this general situation. Now we are posing the problem,

$$P(F_j \,|\mathcal{D}) = P(M_1, M_2, \cdots, M_{16} \,|\, N_1, N_2, \cdots, N_{16})$$

where $\sum_{i=1}^{16} M_i = 4$ and $\sum_{i=1}^{16} N_i = 32$. That is, we want the state of knowledge surrounding future frequency counts in each of the sixteen cells given the known frequency counts in the cells from past data.

---

[1]Sounds like a model mimicking today's educational system where self–esteem is valued far more than an accurate assessment of a student's capabilities.

Figure 14.1: *32 students placed into a contingency table consisting of 16 cells that represent graduation status and results on three tests.*

Let's calculate the probability of seeing all four future students GRADUATE with HIGH scores on all three tests. This stipulates cell 1 of the contingency table, and so we set $M_1 = 4$ with all fifteen remaining $M_i = 0$. The N_i are read off from the contingency table containing the data. The prediction formula was derived in Chapter Twelve, and shown there as Equation (12.2).

Rewriting that equation where now $n = 16$, and explicitly spelling out the constant factor C, we have,

$$P(M_1 = 4, M_2 = 0, \cdots, M_{16} = 0 \mid N_1 = 8, N_2 = 2, \cdots, N_{16} = 5) =$$

$$\frac{M! \, (N + n - 1)!}{N_1! \, N_2! \cdots N_{16}! \, (M + N + n - 1)!} \times \frac{\prod_{i=1}^{16} (M_i + N_i)!}{\prod_{i=1}^{16} M_i!}$$

The details of this calculation are shown in Exercise 14.5.4 and the answer is,

$$P(M_1, M_2, \cdots, M_{16} \mid \mathcal{D}) = 0.001981 = 1.981 \times 10^{-3}$$

In the situation discussed in the last Chapter where we didn't take any data into consideration, or where we simply did not possess any data, all possible models were on the same equal footing. Models that claimed Q_1 close to 1, and the other fifteen Q_i close to 0, had to be considered in the same light as models that claimed Q_{16} was close to 1, and all the other Q_i were close to 0.

Just one measly data point, say, $N_1 = 1$, is enough by itself to deductively rule out the vast number of models that asserted $Q_1 = 0$. This piece of data also provides strong evidence against similar models that assert Q_1 is close to 0. With $N = 32$ data points, a huge swath of potential models have been relegated to ignominy, while others have enjoyed a rise to fame.

The probability calculated above of 0.001981 is a result of readjusting all the models by conditioning on the data. The IP is now in a state of knowledge where its degree of belief that this statement is TRUE is over 100 times larger when compared to the fair model.

$$P(F_j \mid \mathcal{M}_{\text{Fair}}) = W(M)\, Q_1^{M_1} \cdots Q_{16}^{M_{16}} = (1/16)^4 = 0.000015 = 1.50 \times 10^{-5}$$

The data are providing some evidence that the information processor should consider many models that are not the fair model. Thus, the probability that all four students do graduate with HIGH test scores is higher than the probability provided from the fair model because there is some evidence to that effect in the past data. The past data are changing the probability of the models that are used to average over all the predictions about the characteristics of the students.

Shouldn't a state of knowledge about future frequency counts that just happens to follow the pattern of frequency counts in the data produce a higher probability than the one just calculated? If we ask for the probability that four future students graduate, and obtain test scores somewhat like the past students, we might ask for the probability that one student GRADUATEs with all HIGH test scores, $M_1 = 1$, no students GRADUATE with a HIGH LOW HIGH pattern, $M_2 = 0$, and so on. Plugging these new arguments into the predictive formula reveals that the probability for a future student in cells 1, 6, 7, and 16 is,

$$P(M_1 = 1, M_2 = 0, \cdots, M_6 = 1, M_7 = 1, \cdots, M_{16} = 1 \mid \mathcal{D}) = 0.00432 = 4.32 \times 10^{-3}$$

confirming our intuition about this situation with a higher degree of belief.

The actual numerical values for all these probabilities are quite small. When n and M increase, the number of possibilities increase so rapidly that low level events have a very small probability. We would have to sum over very many such events before we reached the realm where the probabilities seemed pragmatically relevant.

14.3   State of Knowledge about the Models

The formal manipulation rules of probability told us that,

$$P(F_j \mid \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(F_j \mid \mathcal{M}_k)\, P(\mathcal{M}_k \mid \mathcal{D})$$

Let's concentrate in this section on the model updating term $P(\mathcal{M}_k \mid \mathcal{D})$ within the overall prediction equation.

The kth model $\mathcal{M}_k$ is a statement assigning "correct" numerical values to $q_1, q_2, \cdots, q_n$, while the data $\mathcal{D}$ are the past frequency counts $N_1, N_2, \cdots, N_n$. The formula for finding the updated probability for some model following the derivation of Chapter Twelve is,

$$P(q_1, q_2, \cdots, q_n \mid N_1, N_2, \cdots, N_n) = \frac{(N + n - 1)!}{N_1!\, N_2! \cdots N_n!}\, q_1^{N_1}\, q_2^{N_2} \cdots q_n^{N_n}$$

For the particulars of this Chapter's problem,

$$P(q_1, q_2, \cdots, q_{16} \mid N_1, N_2, \cdots, N_{16}) = \frac{(32 + 16 - 1)!}{8!\, 2! \cdots 5!}\, q_1^8\, q_2^2 \cdots q_{16}^5$$

However, we have to be careful here. We are not actually calculating the probability, but the *probability density function* because we made the switch from a discrete model space to a continuous model space for the q_i.

For example, the updated probability density function for the fair model, where all $q_i = 1/16$ and conditioned on the known past data, is calculated as,

$$\begin{aligned}
\mathrm{pdf}\,(q_1, q_2, \cdots, q_{16} \mid N_1, N_2, \cdots, N_{16}) &= \frac{47!}{8!\, 2! \cdots 5!}\, (1/16)^8 (1/16)^2 \cdots (1/16)^5 \\[2mm]
&= 4.55 \times 10^{10}
\end{aligned}$$

If another model were to assign numerical values to the q_i closely matching the normed frequencies, then the *probability density function* would be larger. For example, if a new model were to assign $q_1 = 1/4$ matching $N_1/N = 8/32$ and so on, then,

$$\begin{aligned}
\mathrm{pdf}\,(q_1, q_2, \cdots, q_{16} \mid \mathcal{D}) &= \frac{47!}{8!\, 2! \cdots 5!}\, (1/4 - 0.0004)^8 (1/16)^2 \cdots (3/32)^5 \\[2mm]
&= 1.64 \times 10^{18}
\end{aligned}$$

The numerical assignment to q_1 subtracted 0.0004 from 1/4 because 0.0001 was assigned to those four cells where a frequency count was 0. The overall sum of the q_i had to remain at 1.

14.3.1   Models, data, and entropy

As a foreshadowing of important concepts about information entropy to appear in the second and third Volumes, consider the following discussion of model updating in the presence of data.

We mentioned before that what really is of interest is the *relative* standing between any two models. Bayes's Theorem gives us,

$$P(\mathcal{M}_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(\mathcal{D})}$$

The ratio of any two models, call them model $\mathcal{M}_A$ and model $\mathcal{M}_B$, is then,

$$\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} = \frac{P(\mathcal{D} \mid \mathcal{M}_A)\, P(\mathcal{M}_A)}{P(\mathcal{D} \mid \mathcal{M}_B)\, P(\mathcal{M}_B)}$$

By forming the ratio, we have conveniently gotten rid of the denominator $P(\mathcal{D})$.

Revisit the formula for the *probability density function* of a model conditioned on the data when the IP is uninformed,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{(N + n - 1)!}{N_1! \; N_2! \cdots N_n!} \; q_1^{N_1} \; q_2^{N_2} \cdots q_n^{N_n}$$

Since the leading fractional term containing N, n, and the various N_i is not going to change when we form the ratio of any two models, we can focus just on the term containing the different q_i assignments made by the two models,

$$\frac{P(\mathcal{M}_A \,|\, \mathcal{D})}{P(\mathcal{M}_B \,|\, \mathcal{D})} = \frac{P(\mathcal{D} \,|\, \mathcal{M}_A) \, P(\mathcal{M}_A)}{P(\mathcal{D} \,|\, \mathcal{M}_B) \, P(\mathcal{M}_B)} = \frac{Q_{1A}^{N_1} \, Q_{2A}^{N_2}, \cdots, Q_{nA}^{N_{16}}}{Q_{1B}^{N_1} \, Q_{2B}^{N_2}, \cdots, Q_{nB}^{N_{16}}}$$

Remember that the initial ratio of the models, before any data, is equal to 1.

It's instructive to take the logarithmic transform of the model comparison so that we can get to our entropy expressions,

$$\ln \left[\frac{P(\mathcal{M}_A \,|\, \mathcal{D})}{P(\mathcal{M}_B \,|\, \mathcal{D})} \right] \;\; = \;\; \sum_{i=1}^{16} N_i \ln Q_{iA} - \sum_{i=1}^{16} N_i \ln Q_{iB}$$

$$= \;\; \sum_{i=1}^{16} N_i \ln \left[\frac{Q_{iA}}{Q_{iB}} \right]$$

Multiply by N on the right hand side and then compensate for this multiplication by dividing by N to arrive at,

$$\ln \left[\frac{P(\mathcal{M}_A \,|\, \mathcal{D})}{P(\mathcal{M}_B \,|\, \mathcal{D})} \right] \;\; = \;\; N \sum_{i=1}^{16} \frac{N_i}{N} \ln \left[\frac{Q_{iA}}{Q_{iB}} \right]$$

$$= \;\; N \sum_{i=1}^{16} f_i \ln \left[\frac{Q_{iA}}{Q_{iB}} \right]$$

We have used the convenient notation f_i for the normed frequency counts N_i/N. Undo the logarithmic transform to get back to the original ratio,

$$\frac{P(\mathcal{M}_A \,|\, \mathcal{D})}{P(\mathcal{M}_B \,|\, \mathcal{D})} = \exp \left[\, N \times \text{Summation term} \, \right]$$

where the summation term is an entropy expression,

$$\sum_{i=1}^{16} f_i \ln \left[\frac{Q_{iA}}{Q_{iB}} \right]$$

examined in more detail later on in Volumes II and III.

As a numerical check on this result, look at the ratio of the updated probability for the two models where $\mathcal{M}_A$ is the fair model, and $\mathcal{M}_B$ is the model closely

matching the normed frequencies. Working straight from the definition, we first calculated the ratio of posterior probabilities at these two different assignments to the q_i as,

$$\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} = \frac{4.5452 \times 10^{10}}{1.6356 \times 10^{18}}$$

$$= 2.78 \times 10^{-8}$$

This value matches the calculation based on,

$$\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} = \exp\left[N \times \sum_{i=1}^{16} f_i \ln\left(\frac{Q_{iA}}{Q_{iB}}\right) \right]$$

$$= 2.78 \times 10^{-8}$$

Model $\mathcal{M}_A$'s predictions, that is, the fair model's predictions, are hardly counted at all relative to model $\mathcal{M}_B$'s predictions when the averaging takes place within the prediction equation. The fair model simply has very little support from the data.

More importantly and to the point, using $P(\mathcal{M}_k \mid \mathcal{D})$ in the prediction equation overturned the formerly equal status ascribed to each model. When there were no data, we were forced to use just $P(\mathcal{M}_k)$. Bizarre models, which had to be granted equal status to any other model, can now be discarded if the data do not provide any evidence in their favor. Hand in hand, this is the way theoretical and empirical science make progress.

14.4   Predicting What Will Happen to the *Next* Student Given the Test Scores

Now that we have delved somewhat into what $P(\mathcal{M}_k \mid \mathcal{D})$ means for updating a state of knowledge about the current set of models, let's return to the overall prediction equation,

$$P(F_j \mid \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(F_j \mid \mathcal{M}_k)\, P(\mathcal{M}_k \mid \mathcal{D})$$

and use it to predict the graduation status for the very next student given that the student's scores on all three tests are known.

The very next student is the 33rd student, but the phrase *next student* could be said to apply to any student who does not belong to the 32 students in the data base. Until we change the data base by adding more students with known outcomes, we will continue to call all further students the *next student*, and label all such students with the subscript $N + 1$.

Thus, the future event in this case is $F_j \equiv (A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1})$. The prediction updates a state of knowledge about that next student given that the student has already taken the tests. Thus, the test scores obtained by the next student, as well as all of the past data, are known.

$$P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{D}) =$$

$$\sum_{k=1}^{\mathcal{M}} P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_k) \, P(\mathcal{M}_k \mid \mathcal{D})$$

The first term in the summation on the right hand side is solved by Bayes's Theorem.

$$P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_k) =$$

$$\frac{P(A_{N+1}, B_{N+1}, C_{N+1}, D_{N+1} \mid \mathcal{M}_k)}{P(A_{N+1}, B_{N+1}, C_{N+1}, D_{N+1} \mid \mathcal{M}_k) + P(\overline{A}_{N+1}, B_{N+1}, C_{N+1}, D_{N+1} \mid \mathcal{M}_k)}$$

Suppose that we want to update a state of knowledge about whether the next student will GRADUATE after obtaining HIGH scores on all three tests. It is easy to locate the proper cells in the joint probability table to form Bayes's Theorem. (Refer back to Figure 14.1.) The joint statement in the numerator is,

$A_{N+1} =$ GRADUATE and $B_{N+1} =$ HIGH and $C_{N+1} =$ HIGH and $D_{N+1} =$ HIGH

while the additional joint statement needed in the denominator is,

$\overline{A}_{N+1} =$ DOES NOT GRADUATE and $B_{N+1} = C_{N+1} = D_{N+1} =$ HIGH

Thus, the numerical values assigned under model $\mathcal{M}_k$ to cell 1 and cell 9 are inserted into Bayes's Theorem,

$$P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_k) = \frac{Q_1}{Q_1 + Q_9}$$

A model $\mathcal{M}_B$ discussed in the last section was inspired by the desire to adhere very closely to the observed frequencies in the data. Thus, q_1 was assigned a value close to $1/4$, q_2 the value $2/32$, and so forth under such a model. The prediction about graduation for the next student given high test scores from this model is,

$$P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_B) = \frac{Q_1}{Q_1 + Q_9} = \frac{0.2496}{0.2496 + 0.0001} = 0.9996$$

This can be compared with the prediction from the fair model $\mathcal{M}_A$ which is $1/2$. But the updated probability for $\mathcal{M}_A$, after taking account of the observed data, was so low that such a prediction would have no impact whatsoever.

Consider another model, a model $\mathcal{M}_C$ discussed in the exercises, introduced for the sole purpose of comparing more favorably with model $\mathcal{M}_B$. Thus, we seek to

avoid the fate that befell the fair model which assigned all $q_i = 1/16$. This new model's prediction is,

$$P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_C) = \frac{Q_1}{Q_1 + Q_9} = \frac{0.21}{0.21 + 0.01} = 0.9545$$

In an exercise, it is found that,

$$\frac{P(\mathcal{M}_B \mid \mathcal{D})}{P(\mathcal{M}_C \mid \mathcal{D})} \approx 4$$

Restricting ourselves to just these two models by averaging their predictions,

$$P(A_{N+1} \mid B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{D}) = (0.9996 \times 0.80) + (0.9545 \times 0.20) = 0.9906$$

Obviously, there is a strong implication that obtaining high test scores will result in graduation. There will be many models supported by the data that have Bayes's Theorem predicting probabilities close to 1 for graduation after obtaining high test scores. Any model that has Bayes's Theorem predicting a probability like model $\mathcal{M}_A$'s 1/2 will not count for much in the overall prediction equation.

As information processors, we are practically certain that the next student will graduate given a stellar performance on the battery of tests. What is the reason for this high level of certainty?

The relatively high probability of 0.2496 or 0.21 as assigned to cell 1 by models $\mathcal{M}_B$ or $\mathcal{M}_C$ is one factor. The other factor is that the numerical value assigned to cell 9 under the two models, that is, the joint statement that the student does not graduate given high scores on all three tests, at either 0.0001 or 0.01, was so low.

Both assignments were directly motivated by the observed normed frequencies in the past data on $N = 32$ students. The third factor is that we artificially restricted ourselves to consideration of just two models. A more thorough inclusion of models would have brought the predicted probability down somewhat.

In addition, the restricted nature of how the state space was defined plays a significant role. Could it be that there are other causal factors other than those three tests that impact a student's graduation? It seems very likely there exist many other influences. If we don't account for them, then by ignoring them we are injecting some "noise" into the system.

Refer to the final exercise to explore the implications of different data. For example, if the actual data had fewer students graduating with all high test scores, and more students not graduating with high test scores, then we would not be as certain that the next student who did in fact obtain all high scores would graduate. The conclusions change in the direction one would expect.

14.5   Solved Exercises for Chapter Fourteen

Exercise 14.5.1. Think of another example for one of the sixteen possible joint statements of this Chapter.

Solution to Exercise 14.5.1

We are asking for another example from the state space consisting of $n = 16$ joint statements. Generically, a joint statement looks like,

$$(A = a_1 \text{ or } a_2, B = b_1 \text{ or } b_2, C = c_1 \text{ or } c_2, D = d_1 \text{ or } d_2)$$

Thus, another possibility for a joint statement might be,

$$(A = a_1, B = b_1, C = c_2, D = d_1)$$

explicitly verbalized as, "Student GRADUATES with a HIGH score on the first test, a LOW score on the second test, and a HIGH score on the third test." A probability for this joint statement would appear in cell 2 in the joint probability table.

Any one of these 16 statements comprising the state space can in principle, and in fact, be unambiguously judged as TRUE or FALSE. They are also mutually exclusive and exhaustive statements. A student may be characterized by one, and only one, of these properties. Before the student has been observed to fall into one of these 16 categories, the information processor assigns a probability to indicate its state of knowledge about the statement. Every single model considered makes just such an assignment.

Prior to a student taking the tests, everything is unknown. After a student takes the tests, but prior to graduation, we could use Bayes's Theorem, conditioning on the test scores, to update a state of knowledge about graduation. When the student either graduates or drops out, any remaining uncertainty collapses into a statement that is now unambiguously TRUE or FALSE. That student may then be placed into the contingency table as a data point.

Exercise 14.5.2. Describe one of the 65,536 elementary points.

Solution to Exercise 14.5.2

Our four students are named Alex, Beth, Carl, and Dawn. Here is a simple event about graduation and test score characteristics at the lowest level: Alex DOES NOT GRADUATE with LOW, LOW, and HIGH scores. But Carl and Dawn both GRADUATE with HIGH, HIGH, and LOW scores. Beth GRADUATES with HIGH, HIGH, and HIGH scores. This is as detailed a statement as we can get.

Exercise 14.5.3. Summarize the counting for events involving four future students with 16 possible characteristics.

Solution to Exercise 14.5.3

First of all, there are a total of $n^M = 16^4 = 65,536$ elementary points in the sample space to be accounted for. One of these elementary points was just described in the previous exercise. There are a total of,

$$\frac{(M+n-1)!}{M!\,(n-1)!} = \frac{(4+16-1)!}{4!\,15!} = 3,876$$

macro–statements describing possible frequency counts adding up to the required $M = 4$ scattered among the 16 cells of the contingency table.

This sum of 3,876 is decomposed just like the example in the last Chapter since M, the sum of the future frequency counts in the contingency table, is still 4. That is, we look at the decomposition of the sum 4 into its five possible cases with now a lot more 0s to contend with in the sixteen cell contingency table,

Case 1. $4 = 4 +$ fifteen 0s,

Case 2. $4 = 3 + 1 +$ fourteen 0s,

Case 3. $4 = 2 + 2 +$ fourteen 0s,

Case 4. $4 = 2 + 1 + 1 +$ thirteen 0s,

Case 5. $4 = 1 + 1 + 1 + 1 +$ twelve 0s

For example, how many events from the total of 3,876 are there for the fourth case listed above where two students share the same trait, and the others have different traits? The first term in the counting formula gives us,

$$\frac{n!}{r_z!\times r_s!\times r_d!\times\cdots\times r_M!} = \frac{16!}{13!\,2!\,1!\,0!\,0!} = 1,680$$

different ways for this frequency count to happen if we don't happen to care about the particular way the zero, the single, and the double frequency counts happen. The terms in the denominator refer to $r_z = 13$ (thirteen repetitions of a zero count in the sixteen cells), $r_s = 2$ (two repetitions of a single count), and $r_d = 1$ (one repetition of a double count). The two remaining r_i are both equal to 0 because there are no three or four counts.

The easiest case to check intuitively is the first case where all four students congregate into one cell of the contingency table. A mechanical application of the formula tells us that there are sixteen possibilities,

$$\frac{n!}{r_z!\times r_s!\times r_d!\times\cdots\times r_M!} = \frac{16!}{15!\,0!\,0!\,0!\,1!} = 16$$

We confirm this by repeatedly placing all four students into each one of the sixteen cells of the contingency table with 0s placed in the remaining fifteen cells.

If we add up the total number of ways for these frequency counts to occur over the five cases listed above, we must find that they equal 3,876. The first and fourth cases have just been calculated, and when we do the same thing for the second, third, and fifth cases we find that indeed,

$$16 + 240 + 120 + 1680 + 1820 = 3,876$$

If we bring in the multiplicity factor, the second term in the counting formula, then we can count up all the different ways contributing to the overall sum of 65,536 when we take into account the fact that the students are different individuals.

We just calculated that there were 1,680 ways to scatter the frequency counts of 0, 1, and 2 over the sixteen cells of the contingency table where these frequency counts must sum to four. Suppose we pick one of these $1,680$ possibilities to be where there is a single frequency count in cell 6, another single frequency count in cell 16, a double frequency count in cell 10, and zero frequency counts in the other thirteen cells. In other words, we are specifying that,

$$M_1 = 0, M_2 = 0, \cdots, M_6 = 1, \cdots, M_{10} = 2, \cdots, M_{16} = 1 \text{ where } \sum_{i=1} M_i = 4$$

There are 12 different ways the four students might distribute themselves for this particular frequency count in the contingency table as calculated by the multiplicity factor,

$$W(M) = \frac{M!}{M_1! \, M_2! \, \cdots \, M_{16}!} = \frac{4!}{0! \, 0! \, \cdots 2! \, \cdots \, 0! \, \cdots \, 1! \cdots \, 0! \, \cdots \, 1!} = 12$$

Thus, there are $1,680 \times 12 = 20,160$ elementary points for this event.

Table 14.1 at the top of the next page summarizes all of the remaining similar calculations. The sum over all of the elementary points in the sample space does indeed sum up to 65,536.

The fraction of ways where all four students share the same trait is 16/65536. In contrast, the fraction of ways that they all have different traits is 43680/65536. If we define an event where at most two students share a trait, then this fraction jumps up to 64560/65536. Under the "fair" model, it is highly unlikely to ever see three or four students share the same characteristic.

Conversely, if we were ever to observe students clustering together into some cell, then we would take this as evidence against the fair model that a value of 1/16 should be assigned to the joint statements.

Table 14.1: *Counting formula showing how 65,536 elementary points are aggregated to define higher level events.*

Breakdown	First term	Second term	Multiplication
$4+0+0+0$	$\frac{16!}{15!\,0!\,0!\,0!\,1!} = 16$	$\frac{4!}{4!\,0!\,0!\,0!} = 1$	$16 \times 1 = 16$
$3+1+0+0$	$\frac{16!}{14!\,1!\,0!\,1!\,0!} = 240$	$\frac{4!}{3!\,1!\,0!\,0!} = 4$	$240 \times 4 = 960$
$2+2+0+0$	$\frac{16!}{14!\,0!\,2!\,0!\,0!} = 120$	$\frac{4!}{2!\,2!\,0!\,0!} = 6$	$120 \times 6 = 720$
$2+1+1+0$	$\frac{16!}{13!\,2!\,1!\,0!\,0!} = 1680$	$\frac{4!}{2!\,1!\,1!\,0!} = 12$	$1680 \times 12 = 20160$
$1+1+1+1$	$\frac{16!}{12!\,4!\,0!\,0!\,0!} = 1820$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$1820 \times 24 = 43680$
Totals	3876		65536

Exercise 14.5.4 Work out the details for the problem in section 14.2.2 for calculating the probability of seeing four future students graduate with high scores on all three tests.

Solution to Exercise 14.5.4

The total number of future frequency counts is $M = 4$. The joint statement of a student who graduates with high scores on all three tests is indexed by cell 1 of the joint probability table, so we are interested in $M_1 = 4$ with the remaining frequency counts for the other cells all equal to 0.

The data have fixed $N = 32$ with the frequency counts in the contingency table provided by $N_1 = 8$, and so on. The dimension of the state space for all of the joint statements is $n = 16$. With these specifications, the prediction formula can be filled in and solved.

$$P(M_1 = 4, M_2 = 0, \cdots, M_{16} = 0 \mid N_1 = 8, N_2 = 2, \cdots, N_{16} = 5) =$$

$$\frac{M!\,(N+n-1)!}{N_1!\,N_2!\cdots N_{16}!\,(M+N+n-1)!} \times \frac{\prod_{i=1}^{16}(M_i+N_i)!}{\prod_{i=1}^{16}M_i!}$$

$$= \frac{4!\,(32+16-1)!}{8!\,2!\cdots 5!\,(4+32+16-1)!} \times \frac{(4+8)! \times (0+2)! \times \cdots \times (0+5)!}{4! \times 0! \times \cdots \times 0!}$$

$$= \frac{1}{8! \times 2! \times \cdots \times 5! \times 51 \times 50 \times 49 \times 48} \times 12! \times 2! \times \cdots \times 5!$$

$$= 0.001981$$

Exercise 14.5.5 Solve the same kind of problem when all four students have different traits.

Solution to Exercise 14.5.5

We use the same formula; it's just a matter of substituting different values for the M_i. Instead of $M_1 = 4$, four of the M_i will equal 1. Verify that the updated probability mentioned in the last part of section 14.2.2 is really,

$$P(M_1 = 1, M_2 = 0, \cdots, M_6 = 1, M_7 = 1, \cdots, M_{16} = 1 \mid \mathcal{D}) = 0.00432$$

For a change of pace, compute the constant term C and the second term without performing any cancelations of the factorials,

$$P(M_1 = 1, M_2 = 0, \cdots, M_6 = 1, M_7 = 1, \cdots, M_{16} = 1 \mid \mathcal{D}) =$$

$$\frac{M! \, (N + n - 1)!}{N_1! \, N_2! \cdots N_{16}! \, (M + N + n - 1)!} \times \frac{\prod_{i=1}^{16}(M_i + N_i)!}{\prod_{i=1}^{16} M_i!}$$

$$= \frac{4! \, 47!}{8! \, 2! \cdots 5! \, 51!} \times \frac{(1 + 8)! \times (0 + 2)! \times \cdots \times (1 + 5)!}{1! \times 0! \times \cdots \times 1!}$$

$$= C \times \frac{9! \times 2! \times \cdots \times 6!}{1! \times 0! \times \cdots \times 1!}$$

$$= (2.39 \times 10^{-16}) \times (1.81 \times 10^{13})$$

$$= 0.00432$$

Exercise 14.5.6 Devise a scenario where all four students have different traits not supported by the previous data.

Solution to Exercise 14.5.6

This is essentially the same as the previous problem. Just pick four $M_i = 1$ where the contingency table has zero counts. Referring back to the contingency table in Figure 14.1, we see that 0s are located in cells 9, 11, 13, and 15. We can re–use the C term computed in the last exercise. Thus, we are going to calculate,

$$P(M_1 = 0, \cdots, M_9 = 1, \cdots, M_{11} = 1, \cdots, M_{13} = 1, \cdots, M_{15} = 1, M_{16} = 0 \mid \mathcal{D})$$

$$P(M_1, M_2, \cdots, M_{16} \mid N_1, N_2, \cdots, N_{16}) =$$

$$\frac{M! \, (N + n - 1)!}{N_1! \, N_2! \cdots N_{16}! \, (M + N + n - 1)!} \times \frac{\prod_{i=1}^{16}(M_i + N_i)!}{\prod_{i=1}^{16} M_i!}$$

$$= C \times \frac{8! \times \cdots \times 1! \times \cdots \times 1! \times \cdots \times 1! \times \cdots \times 1! \times 5!}{0! \times \cdots \times 1! \times \cdots \times 1! \cdots \times 1! \times \cdots \times 1! \times 0!}$$

$$= (2.39 \times 10^{-16}) \times (1.67 \times 10^{10})$$

$$= 4.00 \times 10^{-6}$$

This scenario where the four students do not graduate after obtaining high scores on one or more tests has an extremely small probability. This is due to the lack of support by the data for models which would predict students falling into these four cells of the contingency table.

Exercise 14.5.7 Is it possible to calculate an even larger value for the probability density function of a model than presented in section 14.3?

Solution to Exercise 14.5.7

Yes. If another model were to assign numerical values to the q_i even more closely matching the normed frequencies, then the probability density function would be larger. Construct a new model where 0.00004 is subtracted from $q_4 = 1/32$ and 0.00001 is assigned to q_9, q_{11}, q_{13}, and q_{15}. Then,

$$\text{pdf}(q_1, q_2, \cdots, q_{16} \,|\, N_1, N_2, \cdots, N_{16}) =$$

$$\frac{47!}{8!\, 2! \cdots 5!} \, (1/4)^8 (1/16)^2 \cdots (1/32 - 0.00004)^1 \cdots 0.00001^0 \cdots (3/32)^5$$

$$= 1.6546 \times 10^{18}$$

Exercise 14.5.8 Show in more detail the calculation of the ratio of the updated model probabilities in section 14.3.1.

Solution to Exercise 14.5.8

The summation term is a sum over $n = 16$ separate terms. The first term is the normed frequency count appearing in cell 1 of the *contingency table*, N_1/N, times the natural log transform of the ratio of the numerical assignments made by each of the two models to cell 1 of the *joint probability table*.

$\mathcal{M}_A$, the fair model, assigns $Q_{1A} = 1/16$ and $\mathcal{M}_B$, the model close to the data, assigns $Q_{1B} = (1/4 - 0.0004)$ to cell 1 of the joint probability table. The value 0.0004 comes about because 0.0001 was arbitrarily assigned to the four cells in the contingency table containing 0s. This assignment to Q_{1B} took care of the implications of the four 0.0001 assignments in order that the overall sum would remain at 1.

$$f_1 \ln \left[\frac{Q_{1A}}{Q_{1B}} \right] \quad = \quad \frac{8}{32} \ln \left[\frac{1/16}{(.25 - 0.0004)} \right]$$

$$= \quad 1/4 \ln (0.2504)$$

$$= \quad -0.3462$$

The second term is the normed frequency count for cell 2, N_2/N, times the natural log transform of the ratio of the numerical assignments made by each of the two models to cell 2 of the joint probability table. $\mathcal{M}_A$, the fair model, still assigns $Q_{2A} = 1/16$ and $\mathcal{M}_B$, the model close to the data, also assigns $Q_{2B} = 1/16$ because $N_2/N = 2/32$.

$$f_2 \ln \left[\frac{Q_{2A}}{Q_{2B}} \right] \quad = \quad \frac{2}{32} \ln \left[\frac{1/16}{1/16} \right]$$

$$= \quad 1/16 \ln (1)$$

$$= \quad 0$$

If the two assignments are the same, then the summation is not increased.

The final term is the normed frequency count for cell 16, N_{16}/N, times the natural log transform of the ratio of the numerical assignments made by each of the two models to cell 16 of the joint probability table. $\mathcal{M}_A$, the fair model, assigns $Q_{16A} = 1/16$ because it assigns the same value to every cell, and $\mathcal{M}_B$, the model close to the data, assigns $Q_{16B} = 5/32$ because $N_{16}/N = 5/32$.

$$f_{16} \ln \left[\frac{Q_{16A}}{Q_{16B}} \right] \quad = \quad \frac{5}{32} \ln \left[\frac{1/16}{5/32} \right]$$

$$= \quad 5/32 \ln (0.4000)$$

$$= \quad -0.1432$$

The sum of all 16 terms calculated in this manner is,

$$\sum_{i=1}^{16} f_i \ln \left[\frac{Q_{iA}}{Q_{iB}} \right] = -0.543708$$

The final step is to multiply this sum by $N = 32$, and undo the logarithmic transform by taking the exponential transform,

$$\frac{P(\mathcal{M}_A \,|\, \mathcal{D})}{P(\mathcal{M}_B \,|\, \mathcal{D})} = \exp \left[\, 32 \times (-0.543708) \, \right] = 2.78 \times 10^{-8}$$

Exercise 14.5.9 Think of another model that would have more of an influence on a prediction than the fair model.

Solution to Exercise 14.5.9

We have just shown that model $\mathcal{M}_A$'s predictions will not count for much in any prediction about a future event. It's not hard to surmise that another model, call it model $\mathcal{M}_C$, that is also fairly close to the data *a la* model $\mathcal{M}_B$ will have more of an impact on a prediction. We want model $\mathcal{M}_C$ to play a more roughly equal role in the prediction process.

We'll select a model $\mathcal{M}_C$ that is not quite as good as model $\mathcal{M}_B$ in that it doesn't match up with the data quite as well, but nonetheless is not overwhelmed by model $\mathcal{M}_B$'s stature as model $\mathcal{M}_A$ was.

Choose $Q_{1C} = 0.21$ and $Q_{9C} = Q_{11C} = Q_{13C} = Q_{15C}$, the cells where the 0s are located in the contingency table, to have a value of 0.01. The rest of the assignments are the same for both models. This assignment for model $\mathcal{M}_C$ balances off the available probability, but does not match the data quite as well as model $\mathcal{M}_B$'s $Q_{1B} = 0.2496$ and $Q_{9B} = Q_{11B} = Q_{13B} = Q_{15B} = 0.0001$.

Directly, we have,

$$\frac{P(\mathcal{M}_B \,|\, \mathcal{D})}{P(\mathcal{M}_C \,|\, \mathcal{D})} = \exp\,[\,N \times \text{Summation term}\,]$$

with

$$\text{Summation term} = \sum_{i=1}^{16} f_i \ln\left[\frac{Q_{iB}}{Q_{iC}}\right]$$

Just like the last exercise, we churn through the summation (all of this, of course, is done by a *Mathematica* program) starting with the first cell,

$$
\begin{aligned}
f_1 \ln\left[\frac{Q_{1B}}{Q_{1C}}\right] &= \frac{8}{32} \ln\left[\frac{0.2496}{0.21}\right] \\[2mm]
&= 1/4 \,\ln\,(1.1886) \\[2mm]
&= 0.0432
\end{aligned}
$$

But the entire summation is encapsulated by the difference in assignments within this first term. All of the other terms are equal to 0 because the assignments from the two models are the same, or because where there is a difference, this is multiplied by $f_i = N_i/N = 0$ since the frequency counts are 0 in these four cells.

$$\frac{P(\mathcal{M}_B \,|\, \mathcal{D})}{P(\mathcal{M}_C \,|\, \mathcal{D})} = \exp\,[\,32 \times 0.0432\,] = 3.98$$

Thus, model $\mathcal{M}_B$'s predictions are weighted about four times those of model $\mathcal{M}_C$. However, now we would want to take into account and weight appropriately whatever model $\mathcal{M}_C$'s prediction might be.

Exercise 14.5.10. What impact would there be on the conclusions at the end of section 14.4 if the data had been slightly different?

Solution to Exercise 14.5.10

Suppose that cell 1 of the contingency table had only 6 students as opposed to 8, and those two students now showed up in cell 9 where previously there were no students. What we are trying to do here is clear. We are reducing the support from the data for the implication that high test scores are related to graduation.

Accordingly, change model $\mathcal{M}_B$'s assignment for Q_{1B} to $0.1875 - 0.0003$ and Q_{9B} to $1/16$. Change model $\mathcal{M}_C$'s assignment for Q_{1C} to $0.1875 - 0.03$ and Q_{9C} also to $1/16$. The new ratio of updated model probabilities as conditioned on the new data is,

$$\frac{P(\mathcal{M}_B \,|\, \mathcal{D})}{P(\mathcal{M}_C \,|\, \mathcal{D})} = 2.819$$

The individual predictions by Bayes's Theorem given the change in the models are,

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_B) = \frac{Q_{1B}}{Q_{1B} + Q_{9B}} = \frac{0.1872}{0.1872 + 0.0625} = 0.7497$$

and,

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{M}_C) = \frac{Q_{1C}}{Q_{1C} + Q_{9C}} = \frac{0.1575}{0.1575 + 0.0625} = 0.7159$$

The numerical assignment to cell 1 of the joint probability table can no longer be as large as before, while the assignment to cell 9 has to be larger than before if we are to keep pace with the changed data.

Again, restricting ourselves to just these two models,

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{D}) = \frac{(0.7497 \times 2.819) + (0.7159 \times 1)}{3.819} = 0.7408$$

As information processors, we are not as certain as before that the next student will graduate given high scores on the battery of tests. The data have shifted their support to new models where the output from Bayes's Theorem indicates less certainty about this outcome.

Instead of just the two models mentioned in this exercise, if the IP were to average over *all* models under the condition of total ignorance about the causes of graduation and test scores, then the posterior predictive probability would be modulated downwards to 0.70.

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{D}) = \frac{N_1 + 1}{(N_1 + 1) + (N_9 + 1)}$$

$$= \frac{6 + 1}{(6 + 1) + (2 + 1)}$$

$$= 0.70$$

Chapter 15

What Does Uninformed Mean?

15.1 Introduction

As promised earlier, this Chapter presents an in–depth discussion of what the word *uninformed* might mean in the context of probability theory. Then we move on to some of the rather startling ramifications for the inferences made by an IP who suffers under the sobriquet of "totally ignorant."

In a text devoted to explaining *information* processing, there would be some merit in exploring a concept like *uninformed*, thus highlighting a core concept by emphasizing, so to speak, the background rather than the foreground. However, I also intend for this Chapter to explain more than this seemingly circumscribed goal might suggest.

We have nearly reached the end of this initial stage introducing some of the core concepts of information processing. Thus, with Volume I drawing to a close, a new exciting stage looms on the horizon with the discussion of the Maximum Entropy Principle in Volume II.

It is time to attempt some sort of summing up. Under the guise of trying to explicate Jaynes's fascinating ideas about the seemingly slippery probability concept "uninformative," we are going to talk about a lot of other things that have made an appearance over the past few pages.

We begin by giving, in complete detail, the solution to the "kangaroo problem" that appeared in a lengthy article by Jaynes. This whimsical example was based on earlier versions given by Gull and Skilling. It is a frustrating fact that Jaynes's paper did not concentrate on the single issue we wish to cover here, but spread itself over many disparate topics.

My main purpose is to focus on Jaynes's analysis surrounding the meaning of "uninformative" in the context of the prior probability for models. With this goal in mind, we will expand his abbreviated and subdued numerical examples into much more detail than he chose to do. We are attempting to emphasize Jaynes's startling conclusions with somewhat more vigor than he employed.

The debate over "uninformative" distributions continues unchecked within the Bayesian community. For example, there appears to be no consensus concerning the basic distribution to be employed when an IP professes total ignorance about the models employed to assign numerical values to probabilities. After reading the extensive literature devoted to this topic, one begins to despair of ever finding a way out of the labyrinth.

However, there is some hope. I believe that Jaynes provided us with the best argument for understanding why particular numerical assignments are made by "prior probabilities for models." Unfortunately, his explanation was done in a round about manner, in a relatively obscure publication, and buried amidst a host of other distracting tangential issues. I have attempted here to unwind his essential argument in a slower, more connected fashion with several numerical examples.

Jaynes set the tone for my discussion with this quote appearing late in his article:

> A major thing to be learned in developing this neglected half of probability theory is that the mere unqualified epithet "uninformative" is meaningless. A distribution which is uninformative about variables in one space need not be uninformative about related variables in some other space.

I believe that this issue is best approached by taking advantage of the pleasant mathematical features that arise when combining the multinomial and Dirichlet distributions. By systematically manipulating the parameters appearing in the Dirichlet distribution, one can easily discern the origin for the various "ignorance priors" that have appeared in the Bayesian literature.

One of my goals is to try to clarify just what these different spaces are that Jaynes is referring to. We begin with definitions of 1) the state space, 2) the sample space, and 3) elementary points. Later, we will expand our comments to include the most important space of all, 4) model space.

We will once again leverage the various combinatorial formulas developed for the classical occupancy problem. My discussion is based on examples provided by Feller in his classic text. Unlike much of my discussion here, Jaynes did not treat the kangaroo problem from the standpoint of a classical occupancy problem.

As part of the excruciating detail needed to explain the kangaroo scenario, the discussion is supplemented with a large number of numerical exercises. Among other things, these exercises contain a complete listing of all possible contingency tables, and elementary points for the specific state space and sampling size dictated by the kangaroo scenario. Appendix E contains an extended commentary on the *Mathematica* program used in the numerical computations of this Chapter.

15.2 The Kangaroo Scenario

15.2.1 Introducing the kangaroos

To begin, let us introduce the kangaroos who will play the role of illustrating our theoretical concepts. Apparently, kangaroos in Australia can belly up to the bar in their pub of choice. Our task is to categorize each and every kangaroo according to which beer it prefers, as well as which hand is used to hoist its favored brew.

To keeps things simple, consider only two brands of beer, and two hands (if we might be allowed this license for whatever is the correct biological term). Thus, every kangaroo will be categorized according to one of four traits. The following notation is used where R stands for right–handed, $\overline{R}$ for left–handed, F for a Foster's beer drinker, and $\overline{F}$ for a Corona drinker.

The state space then consists of statements about the joint occurrence of hand preference and beer preference. The dimension of this state space is $n = 4$. Every kangaroo observed must be placed into one of the following four mutually exclusive and exhaustive categories: (1) right–handed Foster's drinker (RF), (2) right–handed Corona drinker ($R\overline{F}$), (3) left–handed Foster's drinker ($\overline{R}F$), and (4) left–handed Corona drinker ($\overline{R}\,\overline{F}$).

We now pose the inferential problem. What is the probability for some number of kangaroos to possess these four traits given that we have never previously sampled any of the kangaroos? Suppose that we arbitrarily pick, as Jaynes did, 16 kangaroos for the numerical example. Thus, we can now specify that $M = 16$.

This can be thought of as 16 repeated trials, just as in the repeated trials of the toss of a coin, or roll of a die. The coin or the die is physically the same on every trial, but there are uncontrollable or unknowable physical influences that change at every trial. In like manner, every kangaroo is considered to be the same except for uncontrollable or unknowable genetic and environmental factors that dictate hand and beer preference.

Eventually, the IP will want to calculate the probability for observing the number of kangaroos who possess each of the four traits. We write this as, for example,

$$P(M_1 = 6, M_2 = 4, M_3 = 1, M_4 = 5)$$

This represents the IP's state of knowledge for a future occurrence where sixteen kangaroos are observed, six of whom are right–handed Foster's drinkers, four right–handed Corona drinkers, one left–handed Foster's drinker, and five left–handed Corona drinkers. The M_i indicate the future frequency counts for the ith trait, where the sum must equal 16, $\sum_{i=1}^{4} M_i = M = 16$.

Notice especially that the IP is NOT conditioning this inference upon any data. Namely, there are no previous frequency counts, and thus no N_i available.

15.2.2   State space, sample space, and elementary points

The *state space* consists of the four joint statements about hand and beer preference. The dimension of the state space is $n = 4$. To keep the notational clutter to a minimum, these statements are expressed in a short form as,

$$(A = a_1), \cdots, (A = a_4)$$

with any abstract probability for a statement written as $P(A = a_i)$.

Thus, $P(A = a_3)$ stands for the IP's degree of belief that the joint statement, "The kangaroo is a left–handed Foster's drinker." is TRUE. It is important to keep emphasizing that the state space consists of *statements*, not numbers, and that the total probability of 1 is distributed over these four statements.

The numerical example employed by Jaynes, and repeated here, talked about a future sample of $M = 16$ kangaroos categorized according to these four beer–hand preference traits.

The *sample space* consists of over four billion *elementary points*. In fact, there are exactly $n^M = 4^{16} = 4,294,967,296$ such elementary points constituting the sample space. We are dangerously close to a combinatorial explosion with such a huge number, but we can, exercising some diligence, actually list all of the elementary points for this example.

In order to list all of the elementary points, we must be able to distinguish each kangaroo as an individual. We accomplish this by giving each of the sixteen kangaroos a name. For example, the first four kangaroos are named 1) Alex, 2) Beth, 3) Carl, and 4) Dawn. In a concise format and following Feller, the individual kangaroos are called **a, b, c, d, $\cdots$, p**.

Here is an example of one elementary point from the grand total of 4,294,967,296 elementary points that live in the sample space. Nine kangaroos named **a, c, d, f, h, l, m, n, o** are right–handed Foster's drinkers, three kangaroos named **b, e, k** are right–handed Corona drinkers, three kangaroos named **g, i, p** are left–handed Foster's drinkers, and the final kangaroo **j** is a left–handed Corona drinker.

A subscript from 1 to M will be attached to A to indicate the lth kangaroo. Thus, the notation $P(A_7 = a_3)$ indicates a probability for the joint statement that kangaroo **g**, named George, drinks Foster's beer with its left hand. This probability represents a degree of belief by the IP that this statement about George drinking Foster's with his left hand is, in fact, TRUE.

When we transition from the realm of purely abstract probabilities to making legitimate numerical assignments to the probabilities, the notation will indicate that the probability is conditioned on the assumed truth of some *statement*, traditionally called a model $\mathcal{M}_k$. The purpose of such a model is to assign numerical values to the probabilities for all four statements in the state space as dictated by information inserted by the model. This is written in the standard way as $P(A = a_i \mid \mathcal{M}_k)$.

15.2.3 Contingency tables

It is very important to distinguish between joint probability tables and contingency tables. Contingency tables contain the actual frequency counts of kangaroos as categorized by the four traits. Contingency tables therefore are a way of displaying any already observed data, or any potential future data.

Joint probability tables, on the other hand, display the numerical values assigned as probabilities by some model to the four joint statements. Joint probability tables will be discussed in a later section dealing with a numerical example of an assignment to probabilities by a model.

For the present example, we are supposing that the $M = 16$ kangaroos will be observed in the future, and placed into one of the four cells of the contingency table according to their beer and hand preference. For the purposes of this problem, there are no past data available. Therefore, any N_i and N representing some observations already made about the hand and beer preferences of the kangaroos do not exist.

The number of possible configurations of the contingency table is given by the formula,

$$\textbf{contingency tables} \ = \ \frac{(M + n - 1)!}{M! \, (n - 1)!} \tag{15.1}$$

There will be 969 different contingency tables containing 969 different frequency counts for our particular numerical example involving $n = 4$ and $M = 16$.

$$\frac{(M + n - 1)!}{M! \, (n - 1)!} \ = \ \frac{(16 + 4 - 1)!}{16! \, (4 - 1)!}$$

$$= \ \frac{19 \times 18 \times 17}{6}$$

$$= \ 969$$

One of these possible 969 contingency tables is shown below as Figure 15.1. This

Figure 15.1: *An example of a contingency table where 16 kangaroos have been categorized according to hand and beer preference.*

contingency table shows the 16 kangaroos as they might be categorized, sometime in the future, according to beer and hand preference with 9 RF kangaroos, 3 $R\overline{F}$ kangaroos, 3 $\overline{R}F$ kangaroos, and 1 $\overline{R}\,\overline{F}$ kangaroo.

The marginal totals for hand preference, beer preference, and the overall number of kangaroos are part of the contingency table. Contingency tables for the kangaroo scenario will consist of four cells laid out in a two column by two row pattern. The two columns indicate the beer drinking preference, and the two rows the hand preference. Each of the four cells contains a frequency count, (that is, the M_i), of how many kangaroos might possess the particular beer–hand combination in some future set of $M = 16$ observations.

Within the text, we will visually depict a full–blown contingency table like the one shown in Figure 15.1 in compact form as,

$$\boxed{9}\,\boxed{3}\,\boxed{3}\,\boxed{1}$$

Otherwise, when sketching the contingency table as a figure, it is drawn as a 2×2 table with marginal sums and identifying labels.

15.2.4   The classical occupancy problem

The kangaroo scenario can be thought of as just another example of the classical occupancy problem. The generic language used by Feller in the occupancy problem is couched in terms of distributing r distinguishable balls into n distinguishable cells. Here, the n cells are the n joint statements in the state space, and the r balls are the M kangaroos.

The combinatorial formula introduced in Chapter Thirteen will be used once again to calculate all of the elementary points in the sample space.

$$\textbf{Elementary points } = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1!\, M_2!\, \cdots\, M_n!} \quad (15.2)$$

This is not a trivial formula, and it takes some time to get used to it. We will provide many, many numerical examples, in addition to verbal explanations, to help in this familiarization process. See the exercises at the end of this Chapter for all of the details in applying Equation (15.2).

15.3   Contingency and Joint Probability Tables

At this juncture, we continue to examine some typical contingency tables. Then, we juxtapose a joint probability table to clearly demarcate how these two tables differ conceptually.

Figure 15.2 shows another contingency table. If you refer ahead to Exercise 15.7.5, you will see that this contingency table is one of the 12 possible contingency tables following the $9 + 3 + 3 + 1$ occupancy pattern. It certainly represents a

Figure 15.2: *A second example of a contingency table where 16 kangaroos have been categorized according to hand and beer preference.*

different set of frequency counts for the kangaroos when compared to the one shown in Figure 15.1. Although, at some higher level, it still is merely reflecting a pattern where 9 kangaroos share the same trait, three others share the same trait, and so on.

Figure 15.3 shows another contingency table with a clearly recognizable pattern. Here, all 16 kangaroos are evenly spread out over the four traits with 4 kangaroos appearing with each possible beer–hand preference.

Figure 15.3: *A contingency table with all 16 kangaroos evenly distributed over the four traits.*

As we will discuss in much greater detail in Volume II, it would be conceptually wrong in the extreme to characterize this last contingency table as somehow being a *maximum entropy* configuration. Information entropy is defined solely with respect to a *probability distribution.*

I want to emphasize that information entropy is NOT defined with respect to the frequencies in the contingency table. It simply does not make any sense to attribute entropy to a distribution of frequency counts, whether those counts have already been recorded, or whether they are merely imagined potential counts.

Remember that the contingency table represents either some already observed data, or, as in this case, a future possibility as to how the 16 kangaroos *might* be categorized. The fact that the IP is considering a *future* contingency table is the only reason why the IP is allowed to wrap the probability operator around the frequency counts, $P(M_1, M_2, M_3, M_4)$.

Since these observations will occur in the future, the IP is uncertain about any possible contingency table. However, *after* the data have been collected on the kangaroos, uncertainty no longer exists concerning the contingency table. *After* the data are known, everything becomes conditioned on these known results. Then, probability expressions like $P(M_1, M_2, M_3, M_4 \mid N_1, N_2, N_3, N_4)$ will appear. And, having said this, it is then fun to muse on why Bayes's Theorem insists on finding a $P(N_1, N_2, \cdots, N_n)$ in its solution.

Joint probability tables have the same outward appearance as contingency tables since they also consist of four cells laid out in a 2×2 arrangement as shown in Figure 15.4. But each cell now contains a numerical assignment to a probability for the joint statement indexed by that particular cell. That assignment comes about from the information inserted by some model $\mathcal{M}_k$. Later on in Volume II, we shall

	F	**F̄**	
R	.70 Cell 1	.05 Cell 2	3/4
R̄	.05 Cell 3	.20 Cell 4	1/4
	3/4	1/4	1

Figure 15.4: *An example of a joint probability table where numerical values have been assigned to probabilities for the four joint statements of the kangaroo's beer and hand preference under some model.*

see that this particular numerical assignment shown in Figure 15.4 comes about from a model that incorporates a correlation between the traits.

If different information under a different model is desired, then that different model makes a different assignment than what is shown in Figure 15.4. Perhaps the probabilities will take on the definite numerical values, with the notation,

$$Q_1 = 9/16, \ Q_2 = 3/16, \ Q_3 = 3/16, \ \text{and} \ Q_4 = 1/16$$

for the four cells under this new model. Do not be confused by these new numerical assignments because they happen to be identical to the normed frequency counts in one of the possible contingency table shown in Figure 15.1.

It should be clear that, despite the outward visual similarities between a normed contingency table and a joint probability table, we are talking about distinctly separate concepts. On the one hand, the contingency table reflects frequency counts. On the other hand, the joint probability table reflects the numerical assignments to probabilities made by some model.

Frequency counts, normed or otherwise, are unchanging observed facts. They are ontological in nature. Probabilities are changeable entities that depend upon the information in a model. They are epistemological in nature.

15.4 Changing the Dirichlet Parameters

Finally, we reach the heart of the matter. Jaynes investigated what happened when the α_i parameters in the Dirichlet distribution capturing the probabilities for models were manipulated in a systematic fashion. In this way, he caught a glimmer of what a slippery word "uninformative" could turn out to be.

Appropriately, he then warned us of the dangers that might arise when these subtleties were not respected. It goes without saying that those warnings have been largely ignored even by people who should know better. I repeat Jaynes's numerical experiments in more detail, but the general conclusions are the same as those reported by Jaynes.

Here is the computational formula,

$$P(M_1, M_2, \cdots, M_n) = \frac{M! \, \Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \prod_{i=1}^{n} \frac{\Gamma(M_i + \alpha_i)}{M_i! \, \Gamma(\alpha_i)} \tag{15.3}$$

that we developed in section 13.3 as Equation (13.1) for the probability of future frequency counts when no data were involved. The derivation of this predictive formula followed the outline of Jaynes's proof.

See Exercise 15.7.9 for a derivation inspired by Jaynes, but with many of the missing steps filled in. It will be the workhorse for all of the numerical examples to come. A small *Mathematica* program described in Appendix E was written for Equation (15.3), and it provided the results shown in the upcoming tables.

15.4.1   Letting all of the α_i march to ∞

Here we begin the slow process of unwinding Jaynes's macro through a series of numerical examples. We have already seen that when all of the alpha parameters appearing in the probability for the models are set equal to 1, there is a uniform distribution over the model space. The probability for every possible frequency count, as reflected in every possible contingency table, possesses exactly the same value of 1/969.

We will now let the α_i parameters take on values different from all $\alpha_i = 1$. What happens if the α_i parameters march in lockstep away from $\alpha_i = 1$ to an eventual $\alpha_i \to \infty$? Table 15.1 summarizes the relevant facts.

Table 15.1: *The probability for future frequency counts in two selected contingency tables changes as the α_i parameters increase.*

α_1	α_2	α_3	α_4	$\mathcal{A}$	$P(\text{all } M_i = 4)$	$P(M_1 = 16)$	Ratio
1	1	1	1	4	1.03×10^{-3}	1.03×10^{-3}	1
2	2	2	2	8	2.55×10^{-3}	6.93×10^{-5}	37
5	5	5	5	20	5.91×10^{-3}	1.19×10^{-6}	4956
10	10	10	10	40	8.79×10^{-3}	6.87×10^{-8}	127,927
20	20	20	20	80	1.11×10^{-2}	7.36×10^{-9}	1,514,383
35	35	35	35	140	1.25×10^{-2}	2.07×10^{-9}	6,029,591
∞	∞	∞	∞	∞	1.47×10^{-2}	2.33×10^{-10}	63,063,000

Focus in on the probability for two special contingency tables from the overall total of 969 that are available. The first is where the kangaroos are evenly spread out over the four traits, 4 4 4 4 . Thus, we seek,

$$P(M_1 = 4, M_2 = 4, M_3 = 4, M_4 = 4)$$

The second contingency table is where all the kangaroos are crowded into just one trait, say, 16 0 0 0 . Thus, for this second case we seek,

$$P(M_1 = 16, M_2 = 0, M_3 = 0, M_4 = 0)$$

We concluded above that the probabilities for these two contingency tables had the same value of 1/969 when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$. Now let all four α_i increase to 2, then 5, then 10, and so on, eventually up to ∞. What happens to the probability for these two configurations?

The first configuration gradually becomes more and more probable, while the second gradually becomes less and less probable. The final column is the ratio,

$$\frac{P(M_1 = 4, M_2 = 4, M_3 = 4, M_4 = 4)}{P(M_1 = 16, M_2 = 0, M_3 = 0, M_4 = 0)}$$

Finally, at the asymptote, the evenly spread out configuration is some 63 million times more probable than the crowded configuration. All calculations are based on Equation (15.3), and its implementation as a *Mathematica* program.

In fact, the ratio shown in the final column when all four α_i approach ∞ is just the ratio of the multiplicity factors for the two special configurations. There is only one way in which all 16 kangaroos can be right–handed Foster's drinkers. But the calculation of the multiplicity factor,

$$W(M) = \frac{16!}{4!\,4!\,4!\,4!} = 63,063,000$$

illustrates the enormous number of ways existing for four of the kangaroos to be right–handed Foster's drinkers, four right–handed Corona drinkers, four left–handed Foster's drinkers, and four left–handed Corona drinkers.

In other words, when the uniform distribution over models is used with all $\alpha_i = 1$, the impact of the multiplicity factor is nullified. As the α_i increase, the multiplicity factor begins to play more and more of a role. Eventually, when the $\alpha_i \to \infty$, the sheer number of ways something could happen takes over, and the configuration with the largest multiplicity factor has a probability $W(M)$ times the configuration that can happen in only one way.

15.4.2   Letting the α_i march towards 0

We have explored the broad geographical region extending from $\alpha_i = 1$ through $\alpha_i \to \infty$. There were definite consequences for the probabilities assigned to the two special configurations under these models. The multiplicity factor, telling us the number of ways that a configuration could be formed, was nullified for the uniform model where all $\alpha_i = 1$, but as the α_i increased, the multiplicity factor played a larger and larger role until it was eventually the deciding factor.

Now there is another possibility open to us. There is no reason to restrict the α_i from proceeding in the opposite direction. As strange as it may sound, it is also possible to explore the region where the α_i approach 0 starting from 1. For technical reasons, the Dirichlet integral cannot allow the α_i to equal 0. But the numerical values initially assigned to any one of the 969 contingency tables can be examined numerically, just as we did above, when the α_i are permitted to march in unison downwards from 1 to 0.

Table 15.2 contains the assigned probabilities for five selected contingency tables as computed by Equation (15.3). These five configurations consist of the maximum

multiplicity factor configuration in row 1. The four configurations that can happen in only one way appear in the next four rows. The α_i start off at 0.5, and then descend towards 0 in the succeeding four columns. With all $\alpha_i = 0.5$, the numerical values start to change from the equal assignments that every configuration possessed when all the $\alpha_i = 1$.

Table 15.2: *What happens to the probability for five specially selected contingency tables when all the parameters of the Dirichlet distribution approach 0?*

$\alpha_i \to 0$					
Frequencies	0.5	0.1	0.01	0.001	0.0001
4 \ 4 \ 4 \ 4	0.000329	$\approx 10^{-5}$	$\approx 10^{-8}$	$\approx 10^{-11}$	$\approx 10^{-14}$
16 \ 0 \ 0 \ 0	0.008232	0.1020	0.2266	0.2475	0.2498
0 \ 16 \ 0 \ 0	0.008232	0.1020	0.2266	0.2475	0.2498
0 \ 0 \ 16 \ 0	0.008232	0.1020	0.2266	0.2475	0.2498
0 \ 0 \ 0 \ 16	0.008232	0.1020	0.2266	0.2475	0.2498

A curious thing is to be noted in examining these result. As the α_i become progressively smaller, the probability for the one configuration with the maximum multiplicity factor becomes smaller. Meanwhile, the probability for the crowded configurations become larger. By the time the $\alpha_i = 0.0001$, we observe that the four configurations which can happen in only one way have divided nearly all of the available probability among themselves, leaving the merest table scraps for all the other 965 configurations.

So, in this case, when the parameters of the Dirichlet distribution all approach 0, as opposed to the case just looked at where they all approached infinity, the maximum multiplicity factor configuration and the minimum multiplicity factor configurations reverse position. The four minimum configurations now assume the pre-eminent position. The one maximum configuration is relegated to the lowly position at the bottom of the heap!

15.4.3   Only one α_i goes towards 0

What happens if only one of the α_i is allowed to march toward 0, and the remaining three are fixed at some value? To examine this case numerically, let's arbitrarily choose α_1 to approach 0 while $\alpha_2 = \alpha_3 = \alpha_4 = 1$. Table 15.3 shows the outcome in this case.

The first column contains six special contingency tables. The first three rows all contain configurations where $M_1 = 0$, that is, none of the $M = 16$ kangaroos

are right–handed Foster's drinkers. The last three rows contain configurations with $M_1 > 0$; there are some kangaroos of the 16 sampled who are right–handed Foster's drinkers.

Table 15.3: *Six select contingency tables as $\alpha_1 \to 0$, and $\alpha_2 = \alpha_3 = \alpha_4 = 1$.*

$\alpha_1 \to 0$ while $\alpha_2 = \alpha_3 = \alpha_4 = 1$					
Frequencies	0.5	0.1	0.01	0.001	0.0001
0 16 0 0	2.5×10^{-3}	5.4×10^{-3}	6.4×10^{-3}	6.5×10^{-3}	6.5×10^{-3}
0 5 5 6	2.5×10^{-3}	5.4×10^{-3}	6.4×10^{-3}	6.5×10^{-3}	6.5×10^{-3}
0 9 3 4	2.5×10^{-3}	5.4×10^{-3}	6.4×10^{-3}	6.5×10^{-3}	6.5×10^{-3}
4 4 4 4	6.9×10^{-4}	1.6×10^{-4}	1.6×10^{-5}	1.6×10^{-5}	1.6×10^{-7}
10 3 2 1	4.4×10^{-4}	7.1×10^{-5}	6.6×10^{-6}	6.5×10^{-7}	6.5×10^{-8}
16 0 0 0	3.5×10^{-4}	4.6×10^{-5}	4.1×10^{-6}	4.1×10^{-7}	4.1×10^{-8}

As $\alpha_1 \to 0$, the assignments for the last three rows become smaller and smaller. However, for the first three rows as $\alpha_1 \to 0$, and α_2, α_3, and α_4 remain fixed at 1, the probability grows larger and approaches an asymptote. The assignment for *all* the configurations that contain *any* number of right–handed Foster's drinkers is going to zero, while *any* configuration that contains *no* right–handed Foster's drinkers is becoming larger and equally probable.

It's clear that in these states of knowledge about the models, no right–handed Foster's drinking kangaroos are preferred. At the same time, while the frequency count for M_1 is fixed at 0, any number of kangaroos possessing the remaining three traits are accorded equal status. The multiplicity factor plays very little role for the remaining three traits even when M_1 is not zero.

15.4.4 One α_i approaches 0 and the rest approach ∞

Based on these findings, it might be conjectured that, say, if $\alpha_3 \to 0$ and α_1, α_2, and $\alpha_4 \to \infty$, then frequency counts with no left–handed Foster's drinkers are preferred, and the remaining traits are favored to the extent that they produce a large multiplicity factor. Let's examine this conjecture numerically in Table 15.4.

As we might have surmised from our earlier experience with Table 15.3, the probability for any configuration with $M_3 > 0$ becomes very small. The last three rows contain representative configurations where $M_3 > 0$. But the relative small-ness within this group with $M_3 > 0$ is modulated by the multiplicity factor. The probability for the most spread out configuration in row 4 is much larger than the crowded configuration in row 6. A configuration in this group with a multiplicity

Table 15.4: *Six select contingency tables when* $\alpha_3 = 0.0001$ *and* $\alpha_1 = \alpha_2 = \alpha_4 \to \infty$.

$\alpha_3 = 0.0001$ *while* α_1, α_2 *and* $\alpha_4 \to \infty$				
Frequencies	10	20	30	40
5 5 0 6	3.10×10^{-2}	3.74×10^{-2}	4.01×10^{-2}	4.16×10^{-2}
9 2 0 5	8.28×10^{-3}	7.21×10^{-3}	6.74×10^{-3}	6.47×10^{-3}
16 0 0 0	3.16×10^{-6}	4.75×10^{-7}	2.07×10^{-7}	1.29×10^{-7}
4 4 4 4	1.41×10^{-8}	2.03×10^{-8}	5.47×10^{-10}	2.04×10^{-10}
7 3 5 1	7.79×10^{-10}	4.74×10^{-11}	7.94×10^{-12}	2.13×10^{-12}
0 0 16 0	9.67×10^{-18}	7.31×10^{-22}	2.00×10^{-24}	2.71×10^{-26}

factor between the minimum and the maximum falls between them in the assigned probability as well.

The probability for any configuration with $M_3 = 0$ is larger, as exhibited by the first three rows, but happens to be modulated by the multiplicity factor reflected in the frequency counts in the remaining three cells. A configuration like that in the first row with a large multiplicity factor is favored relative to a configuration like that in the third row with the smallest possible multiplicity factor. As expected, the configuration depicted in the second row falls in between these high and low multiplicity configurations.

15.4.5   One α_i approaches ∞ while the rest are fixed at 1

For the final numerical exploration, consider the case where one of the α_i is allowed to approach infinity while the other parameters remain fixed at 1. Arbitrarily, let $\alpha_2 \to \infty$ while $\alpha_1 = \alpha_3 = \alpha_4 = 1$. Table 15.5 exhibits some relevant patterns.

The first column once again shows six representative configurations from the total of 969 possible contingency tables. Row 1 shows one of the four configurations where things can happen in only one way. All of the kangaroos are right–handed Corona drinkers. In row 3, we have another such configuration where now all the kangaroos are left–handed Foster's drinkers. In row 2, we have specified a configuration that is almost the same as the minimum multiplicity configuration of row 1.

Row 5 is the familiar maximum multiplicity factor configuration where four kangaroos possess each trait. Row 6 is a configuration with no right–handed Corona drinking kangaroos, but otherwise is a configuration with a high multiplicity factor. Rounding out this collection is a configuration in row 4 with an intermediate number of right–handed Corona drinking kangaroos.

Table 15.5: *Six select contingency tables when $\alpha_2 \to \infty$ while $\alpha_1 = \alpha_3 = \alpha_4 = 1$.*

$\alpha_2 \to \infty$ while $\alpha_1 = \alpha_3 = \alpha_4 = 1$				
Frequencies	10	50	100	150
0 16 0 0 0 15 1 0 0 0 16 0	0.067 0.043 3.29×10^{-8}	0.441 0.109 6.80×10^{-16}	0.643 0.090 4.29×10^{-20}	0.739 0.072 1.09×10^{-22}
3 9 2 2 4 4 4 4 5 0 5 6	1.60×10^{-3} 2.35×10^{-5} 3.29×10^{-8}	7.25×10^{-6} 1.99×10^{-10} 6.80×10^{-16}	1.68×10^{-7} 1.90×10^{-13} 4.29×10^{-20}	1.46×10^{-8} 2.38×10^{-15} 1.09×10^{-22}

The next four columns chart the change in probabilities for the configurations as α_2 gets larger and larger while α_1, α_3, and α_4 remain fixed at 1. α_2 is the parameter in the Dirichlet distribution that would be expected to affect events impacting cell 2, right–handed Corona drinking kangaroos.

The probability for configuration 1 with all 16 right–handed Corona drinkers seems to be approaching 1. The probability for the other configurations is getting smaller, corresponding to the number of right–handed Corona drinkers. Even the configuration in the second row is losing out as α_2 increases, albeit at a slower rate. It differs, though, by only one right–handed Corona drinking kangaroo from the contingency table in row 1 that can happen in only one way.

Note that this pattern differs from the pattern discovered when all $\alpha_i \to 0$ and shown in Table 15.2. There, all *four* minimum multiplicity contingency tables split up the available probability as the α_i went to 0. Here, only the *one* minimum multiplicity contingency table with all kangaroos possessing the trait pointed to by the α_i that is approaching infinity gobbles up the probability.

15.5   A Better Appreciation of "Uninformed"

I have presented the fundamental rationale, borrowed in its essentials from Jaynes, trying to shed some light on the meaning of the various "ignorance priors" that appear prominently in the Bayesian literature. The technical apparatus that best addresses this issue is the "Multinomial–Dirichlet" predictive distribution. The prior probability for the models is handled by the Dirichlet distribution, while the frequency counts under any specific model $\mathcal{M}_k$ are seen to follow the Multinomial distribution.

When one systematically manipulates the α_i parameters within the Dirichlet distribution for the prior probability, one can easily discern the origin for the various "ignorance priors." These are affixed with the names of Bayes–Laplace ($\alpha_i = 1$), Jeffreys ($\alpha = 1/2$), Haldane ($\alpha_i \to 0$), and Jaynes ($\alpha_i \to \infty$).

In an attempt to sum up these remarks, it is appropriate to listen to what Jaynes has to say. Among other things, Jaynes is interested in what happens to the influence of the multiplicity factor $W(M)$ as the α_i are manipulated. He is of the opinion that there ought to be a "smooth transition" from the condition where the multiplicity factor reigns supreme to other conditions that are quite immune to the impact of the multiplicity factor.

He finds (as our numerical investigations confirmed) that it is in the region of all proportionally large α_i, where this smooth transition can begin to take place. [13, pg. 45]

> In the inference called for, relative multiplicities are cogent factors. We expect them to be moderated somewhat by the knowledge that kangaroos are a homogeneous species; but surely multiplicities must still retain much of their cogency. Common sense tells us that there should be a smooth, continuous change in our results starting from the pure monkey case to a more realistic one as we allow the possibility of stronger and stronger correlations. Instead, [the $\alpha_i = 1$ case] represents a discontinuous jump to the opposite extreme which denies entropy any role at all in the prior probability. [The case where the α_i approach zero] goes even further and violently reverses the entropy judgments, placing all prior probability on the situations of zero entropy.

What Jaynes labels as the "pure monkey case" is where the probability for a frequency count is proportional to $W(M)$, the multiplicity factor. Thus, models assigning values close to all $Q_i = 1/4$ are strongly favored. We expect to see frequency distributions where the kangaroos are evenly split over all four traits.

There are no correlations between traits under such models. As we saw in our numerical investigation in Table 15.1, if we let the α_i move down in lockstep from large values, then we move away from models favoring independence between the traits to considering models with an increasingly stronger relationship linking the hand and beer preferences.

For example, when the $\alpha_i = 2$, the ratio of seeing an even split for all the kangaroos is only 37 times more probable than seeing every single kangaroo with the same traits. As the α_i parameters approach 1, models exhibiting extremely strong correlations, as well as those with very little correlation between hand preference and beer drinking preference, are all accorded the same status.

Jaynes then turns his attention to the situation of letting all the α_i tend to zero, which, on some views, was held to be "uninformative." We presented a small empirical experiment letting all the $\alpha_i \to 0$. This turns out to be the flip side to letting the α_i grow proportionally larger.

As before, special configurations garner most of the probability, but now these are the configurations where all of the kangaroos have the same trait. Jaynes then forces us to acknowledge that the concept "uninformative" may not be quite as simple as first thought.

There seems to be some sort of reciprocal arrangement going on. You might consider something rather uninformative from one perspective, but from another perspective, it morphs into something quite the opposite, providing information on some different aspect. [13, pg. 45]

> In what sense, then, can we consider small values of α_i to be "uninformative"? In view of [the results outlined here], they are certainly not uninformative about the [frequency counts]. . . .
>
> Our present problem exhibits a similar "uncertainty relation." The monkey multiplicity factor is completely uninformative on the sample space S of n^M possibilities. But on the [model space of the q_i] it corresponds to an infinitely sharply peaked generating function G, a product of delta functions $\delta(q_i - 1/n)$. Conversely, small values of α_i are uninformative about the q_i, but highly informative about the different points in S, in the limit tying the [frequency counts] rigidly together.

The "uncertainty relation" that Jaynes refers to is an analogy to the Heisenberg Uncertainty Principle of quantum mechanics where the more precisely you attempt to measure, say, the position of a subatomic object, the less precisely you can measure its momentum, and *vice versa*.

Analogously, the more "uninformed" we tried to become on the sample space, the more informed we became about the model space. The more "uninformed" we tried to become about the model space, the more informed we became about the sample space.

Two tables were created in an attempt to summarize these uncertainty relations that Jaynes is talking about. In the first table, Table 15.6 appearing on the next page, we comment on the relationship among the model space, the sample space, the multiplicity factor, and the actual calculated probabilities for two very prominent contingency tables.

As the α_i parameters move in lockstep towards infinity, everything becomes concentrated on one model, the fair model. As Jaynes told us, the probability becomes peaked around the one model assigning the numerical values of $1/n$ for the probability for each of the n joint statements in the state space. When the integration in the Multinomial–Dirichlet predictive equation involves this Dirac δ-function representing the probability for all the models, out comes the Multinomial distribution with the specific value of $Q_i = 1/n$,

$$P(M_1, \cdots, M_n) = W(M)\,(1/n)^{M_1} \times (1/n)^{M_2} \times \cdots \times (1/n)^{M_n} = \frac{W(M)}{n^M}$$

The total number of points in the sample space, n^M, appears in the denominator, and the multiplicity factor, $W(M)$, appear in the numerator for the probability of future frequency counts.

Table 15.6: *Summary table with Jaynes's conclusions about uncertainty relationships among the model space, sample space, multiplicity factor, and probability of future occurrences.*

Model Space	Sample Space	Multiplicity	Probability
$\alpha_i \to \infty$ No uncertainty	Maximum uncertainty	Primary role	$P(\boxed{4\,4\,4\,4}) \gg$ $P(\boxed{16\,0\,0\,0})$
$1 < \alpha_i < \infty$ Increasing uncertainty	Decreasing uncertainty	Lesser role	$P(\boxed{4\,4\,4\,4}) >$ $P(\boxed{16\,0\,0\,0})$
$\alpha_i = 1$ Maximum uncertainty	Decreasing uncertainty	No role	$P(\boxed{16\,0\,0\,0}) =$ $P(\boxed{4\,4\,4\,4})$
$\alpha_i = 1/2$ Increasing uncertainty about which deterministic model	Decreasing uncertainty	Roles starting to reverse	$P(\boxed{16\,0\,0\,0}) >$ $P(\boxed{4\,4\,4\,4})$
$\alpha_i \to 0$ Maximum uncertainty about which deterministic model	No uncertainty	Roles are completely reversed	$P(\boxed{16\,0\,0\,0}) \gg$ $P(\boxed{4\,4\,4\,4})$

Although the IP is completely certain about the model in the model space when all the $\alpha_i \to \infty$, all of the elementary sample points are treated the same. Everything is up to the aggregation carried out by the multiplicity factor. The probability of a spread out contingency table like $\boxed{4\,4\,4\,4}$ is vastly greater than the probability of a crowded contingency table like $\boxed{16\,0\,0\,0}$ simply because the former can happen in so many more ways. The one model that has garnered all of the probability exhibits no correlation between hand and beer preference.

As the α_i parameters are allowed to move downwards from these very large values towards 1, the "smooth and continuous transition" that Jaynes is looking for takes place. More and more models in addition to the fair model are accorded some status in determining the average over model predictions. We are no longer quite as certain about the model space.

The multiplicity factor lessens in importance. Some elementary sample points become more important than before because they have not all been leveled out by the multiplicity factor. The probability of our select contingency table | 4 | 4 | 4 | 4 | is being reduced when compared to the probability of that other landmark contingency table | 16 | 0 | 0 | 0 |. Some of these models now being considered contain correlations to varying degrees between hand and beer preference.

And now we have arrived at that very controversial place where all the $\alpha_i = 1$. The IP is "completely ignorant" about the model space. All models are treated as equal.

The multiplicity factor has been nullified. A contingency table like | 4 | 4 | 4 | 4 | that can happen in over sixty three million ways is accorded the same probability as a contingency table like | 16 | 0 | 0 | 0 | that can come about in only one way. Thus, over sixty three million elementary sample points are considered the equal of just one elementary sample point.

Apparently, ignorance about models has not carried over into equal ignorance about sample points. Thus, a completely uninformed IP has to entertain every single possible correlation concerning the kangaroos's traits from no relationship whatsoever to an absolute unbending linkage between the traits.

Think of the territory between $\alpha_i \to \infty$ and $\alpha_i = 1$ as somehow the mirror image of the territory between $\alpha_i = 1$ and $\alpha_i \to 0$. As we journey from $\alpha_i = 1$ down to $\alpha_i = 1/2$ we discern the beginning of a role reversal. Now, the four contingency tables with all 16 kangaroos sharing the same trait are starting to pick up in probability, while the spread out frequency counts are starting to decline in probability.

This has the consequence that the IP must be losing some of its former "total ignorance" about the model space prevalent in the $\alpha_i = 1$ territory. Apparently, some models advocating a strong linkage between hand and beer preference are preferred.

Continue downwards to the asymptotic value $\alpha_i \to 0$, where this glimmer of a trend strengthens. The contingency tables that reflect frequency counts where all of the kangaroos are huddled together in one cell are highly probable. They become so highly probable that *all* of the remaining contingency tables are approaching a probability of 0! These very few elementary sample points out of the vast number inhabiting the total sample space have acquired an importance far beyond their meager multiplicity factors.

The few models postulating a complete causal linkage between traits are the only models that count any more. However, the slippery notion of "uninformed" now insinuates itself back into this seemingly quite informed milieu in a novel way.

The IP is now "completely uninformed" about these few models that postulate strong causality. Thus, for example, the very informative model that assigns a numerical value of 1 for the probability that all kangaroos are right–handed Foster's

drinking kangaroos has the very same standing as the equally informative model that assigns a numerical value of 1 for the probability that all kangaroos are right–handed Corona drinking kangaroos. Likewise for the other two highly informative models that assert a complete causal linkage between beer and hand preference.

So, even though strong causality is asserted in the most undeniable fashion in this situation, the IP is completely uncertain as to where that strong causality manifests itself! This is truly an almost humorous predicament.

A second table, Table 15.7, tries to package up these remarks about the sample space and model space in a slightly different way. The third column associates the historically prominent name to the parameter values in the Dirichlet distribution. The last column highlights the visual appearance of the Dirichlet distribution under different settings of its parameters.

Of course, by placing Jaynes in the first row, I only mean to acknowledge the conventional association of his name with some vague notion of "maximum entropy." Obviously, Jaynes was aware of the subtle distinctions attached to disambiguating the word "uninformative."

Table 15.7: *Another summary version of Jaynes's conclusions about the meaning of* "*uninformative.*"

Model Space	Sample Space	Associated Name	Dirichlet
$\alpha_i \to \infty$ Very informative	Uninformative	Jaynes	Dirac δ function one strong peak
$1 < \alpha_i < \infty$ In–between compromise	In–between compromise		In–between one gentle peak
$\alpha_i = 1$ Uninformative	Informative	Bayes, Laplace Blower	Uniform flat
$\alpha_i = 1/2$ In–between compromise	In–between compromise	Jeffreys	U-shaped gentle peak**s**
$\alpha_i \to 0$ Uninformative about which linkage	Very informative	Haldane	Strongly U-shaped strong peak**s**

Historically, it became quite fashionable to disparage the Bayes–Laplace concept of "uninformed." Even today, one must be prepared to suffer the rebuke of the truly uninformed if one should have the temerity to defend their views. But in the light of the current discussion, in the end it seems that they may have been right after all.

If one understands that by "uninformed" they meant an information processor who had no knowledge of the correct *cause* of some phenomenon, all is clear. As a consequence, the IP must give equal credence to every single model that makes a legitimate numerical assignment to the probabilities for the statements in the state space. To do otherwise would be to admit that the IP does, in fact, know something about the causal structure of the phenomenon. And, as soon as the IP acknowledges this, it disembarks from that ship sailing under the banner of "totally ignorant."

As has been well documented by historians of probability, Sir Harold Jeffreys also thought something was fishy with the Bayes–Laplace view. He couldn't quite reconcile himself on how to incorporate the "scientific mind set of causality" into this overall structure.[1] In other words, he asked himself, if two chemicals combined to form some compound on the first trial, would he just raise the probability to 2/3 that the same reaction would take place the next time? Of course not, he reasoned.

Having seen the sun rise and set, and understanding the geophysical reasons behind it, one knows with certainty the sun will rise again tomorrow morning. Jeffreys reasoned that it would be ludicrous to calculate the probability the sun will rise based on the past frequency counts as Laplace had done.

And so, through a complicated rationalization, he reached the conclusion that the α_i parameters should be set to a value of 1/2 to reflect a state of ignorance. As we saw in our numerical examples, the probability for models that posit a strong causal linkage is raised, and the probability for frequency counts with all kangaroos having the same traits is raised as well.

It seems Jeffreys was struggling to find some compromise between fundamentally incompatible notions. And now with a detailed discussion of Jaynes's exposition in this Chapter, we know what those irreconcilable notions were. He was seeking some sort of compromise between information existing in two different spaces, the model space and the sample space.

But if there is some kind of "Heisenberg Uncertainty Principle" operating on the information between these two spaces, whatever you do to become more *informed* in one space, you are bound by that action to become more *uninformed* in the other space!

At least, we have to give J.B.S. Haldane the courage of his convictions. If you are in for a penny, then you are in for a pound. Why stop, as Jeffreys did, halfway to the goal? If you proceed all the way down to the bottom, then you really are an advocate for strong causality. All kangaroos must share the same beer–hand preference trait.

---

[1] Jeffreys was influenced by Karl Pearson in this regard.

Thus, which beer a kangaroo prefers is directly linked via a physical law to hand preference, just like hydrogen combines with oxygen the same way every time according to physical law. The ironic thing was, as we alluded to above, having pitched your tent squarely in favor of causality, you were forced by the uncertainty principle back to some fundamental ignorance about exactly which strong causal model was, in fact, the right one.

It is clear what Haldane was driving at in his definition of being "uninformed." The world is governed by causes, and if we could just ferret out all of these causal factors, then the IP could deduce effects with certainty. All of the models that assert deductive certainty, and only those kind of models, must be entertained on an equal basis.[2] It seems to me that Haldane would have loved Wolfram's CA, except that he would have been maximally uncertain about which CA was running the world.

It is just like entertaining the four logic functions we investigated in Exercise 13.8.14. The functional assignment was TRUE for only one combination of variables. After specifying the variables, we definitely knew what was false and what was true. The only problem was that we didn't know which of these four logic functions for two variables was the right one.

15.6   Connections to the Literature

Read Bernardo and Smith [2], or Kass and Wasserman [15] as a sample of the voluminous literature on "ignorance priors," and prior probability for models in general. Prepare yourself to be just as confused when you leave as when you entered.

My discussion of the combinatorial counting formula for the classical occupancy problem is based on examples provided by Feller [4]. As mentioned in a footnote, understanding Feller's computations concerning the distribution of accidents during a week becomes easier if you have managed to follow my very discursive explanation and examples. See my supplemental exercises for the complete picture.

The following comments may serve as a bit of contextual background to anyone wanting to tackle Jaynes's original presentation [13] in **Monkeys, Kangaroos, and N**, henceforth abbreviated to MKN. I provide here some helpful tips from someone who has read, and labored over this paper more times than I care to admit.

The major disagreement I have with Jaynes in this paper centers on insufficient clarification about how the numerical assignments are made under various models. It also remains a mystery to me why he, of all people, emphasized frequency counts as indicating information in models! But that discussion will have to wait until Volume II because it depends on details of the Maximum Entropy Principle.

---

[2]I wonder why Haldane didn't support the uniform prior over model space since this is an operational definition of a "universe queerer than we suppose."

The paper seems to have been written about 1984 when Jaynes was spending a sabbatical at St. John's College and the Cavendish Laboratory at Cambridge University, England. The article finally appeared in the 1986 Proceedings of the Sixth annual workshop on Bayesian and Maximum Entropy issues.

The tone of the paper seems to be a reaction, or response, to some discussions or debate he had with John Skilling on the use of maximum entropy in image restoration. Skilling and colleagues had then recently been involved in developing some novel approaches to image processing of astronomical data using the maximum entropy formalism. Skilling had, in fact, specifically invited Jaynes to spend his sabbatical working on these issues with him and his (Skilling's) graduate students in the UK.

As mentioned at the outset, this paper is, in reality, a terrible mish–mash of all sorts of different topics all jumbled together in over thirty pages. The first part is about the singular value decomposition of the point spread function in optics, and its possible role in a new understanding of image pixel data. These ideas play no further role in the rest of the paper. Therefore, it is necessary to skip over much of the content until the kangaroos make their entrance.

The kangaroos, borrowed from a previous example by Gull and Skilling, are introduced on page 36 in the context of a discussion of a rationale for maximum entropy.[3] Jaynes, at long last, after much hemming and hawing in the next few pages, finally begins his substantive discussion of the prior predictive equation on page 40 of MKN.

In these next few pages, Jaynes gets to the heart of the matter. He develops the arguments and equations culminating in his Equation (31), equivalent to my version in Equation (15.3).

Jaynes uses the notation N and N_i for the future frequency counts where I use M and M_i. I have already adopted N and N_i for the observed frequency counts from past data. Jaynes uses the notation K and k_i where I used $\mathcal{A}$ and α_i.

Both Jaynes and I use n for the dimension of the state space, or equivalently, the number of cells in the joint probability table. Jaynes calls his Equation (31) the "joint predictive prior probability." For me, it is not any kind of "prior probability" but a predictive probability for future frequency counts prior to any data.

I essentially followed Jaynes's derivation beginning at his page 40 in my own presentation here, although I have added many numerical examples, changed some of the notation, skipped some of Jaynes's internal justifications, and provided more explanation. This, after all, is a major *raison d'être* of my books.

Make no mistake, though; I did not independently conceive of how to do this. I absorbed it all from trying to decipher Jaynes. In addition to MKN, one can glean insights into Jaynes's thought processes on this topic from Chapter 18 of his book.

---

[3]There is a quite evident typographical or calculational error at the bottom of page 35.

Earlier, during the course of the derivation, Jaynes presented his Equation (23) which he calls a "standard relation of probability theory."

$$p(N_1 \cdots N_n \,|\, \mathcal{I}) = \int \mathrm{d}^n x \; p(N_1 \cdots N_n \,|\, x_1 \cdots x_n) \; p(x_1 \cdots x_n \,|\, \mathcal{I})$$

This is his version of the "standard relation of probability theory," the **Sum** and **Product** rules, for what I wrote as,

$$P(M_1, M_2, \cdots, M_n) =$$

$$\int \cdots \int_{\sum q_i = 1} P(M_1, M_2, \cdots, M_n \,|\, q_1, q_2, \cdots, q_n) \, \mathrm{pdf}\,(q_1, q_2, \cdots, q_n) \, \mathrm{d}q_i$$

Here we see an interesting point. My q_i are interpreted straightforwardly as numerical assignments to probabilities as made by some model. Jaynes uses the x_i in exactly the same way, but chooses to explain things in a rather convoluted style: [13, pg. 41]

> This suggests that we interpret $[\, p(x_1 \cdots x_n \,|\, \mathcal{I}) \,]$ as the prior probability density of those parameters, following from some information $\mathcal{I}$. Note, to head off a common misconception, that this is in no way to introduce a "probability of a probability." It is simply convenient to index our hypotheses by parameters x_i chosen to be numerically equal to the probabilities assigned by those hypotheses;

Jaynes was forced to call his x_i "parameters" to avoid falling into a trap of setting up the infamous "probability of a probability." But, in fact, if one recognizes the simple conceptual notion that model $\mathcal{M}_k$ is a *statement* assigning numerical values to probabilities, then a probability for such a statement, written as,

$$P(\mathcal{M}_k) \equiv \mathrm{pdf}\,(q_1, q_2, \cdots, q_n)$$

is no cause for alarm.

While in all likelihood another book could be written on this topic alone, I will simply put the cat among the pigeons, and assert that Bruno de Finetti's representation theorem is just Jaynes's prior predictive formula in another guise.

What in most contexts is treated opaquely in terms of "mixture of distributions" and "exchangeable variables" is really nothing more than another straightforward application of the formal sum and product manipulation rules.

Furthermore, everything becomes much clearer if we stick to the idea that the "representation theorem" is simply the way that the probability of frequency counts *must be* expressed as the average of the frequency counts under some model with respect to the probability for these models.

15.7   Solved Exercises for Chapter Fifteen

Exercise 15.7.1. Start thinking about the combinatorial formulas and how they can be applied to the kangaroo scenario.

Solution to Exercise 15.7.1

This exercise has the task of forcing us to start thinking about the occupancy patterns, the number of contingency tables, and the elementary points constituting the sample space for the kangaroo scenario. These are all strictly combinatorial problems, and they can be discussed without ever referring to probability.

We begin by revisiting the basic counting formulas as they will now be applied to the kangaroo scenario. We have repeatedly mentioned the enormous number of elementary points for the kangaroo problem. The total number of elementary points in the sample space is,

$$n^M = 4^{16} = 4,294,967,296$$

As a first illustration, how many elementary points out of the more than four billion elementary points in the sample space are contributed by the particular contingency tables where fifteen kangaroos share the same traits? The occupancy number pattern is a decomposition of the sum 16 into $15 + 1 + 0 + 0$.

In how many ways could the sum $M = \sum_{i=1}^{4} M_i = 16$ be formed when fifteen kangaroos are placed into one cell of the contingency table? As a consequence, the single remaining kangaroo appears in another cell of the contingency table, and, of course, no kangaroos appear in the two remaining cells of the contingency table.

From the combinatorial point of view, we can approach the question of the number of contingency tables by asking for the integer partition of 16 into four parts. $16 + 0 + 0 + 0 = 16$ is the first such partition one would think of, and $15 + 1 + 0 + 0 = 16$ is the next one.

Now there are 12 ways of forming a sum such as $15 + 1 + 0 + 0 = 16$. If you imagine a four cell contingency table laid out in the typical 2×2 fashion, there are six different ways where the two 0s could be placed: 1) in the top row, 2) in the bottom row, 3) in the first column, 4) in the second column, or 5), diagonally in cells 1 and 4, or 6) diagonally in cells 2 and 3. For each of these six ways of distributing the two 0s, there would be two ways to flip flop the frequency counts of 15 and 1. Therefore, in total, there are 12 ways to form an occupancy pattern like $15 + 1 + 0 + 0$.

For an example of one of these twelve possible contingency tables recording future frequency counts with the desired occupancy pattern, consider case 5 listed above where the zero frequency counts occur diagonally in cells 1 and 4 of the contingency table. This is illustrated as $\boxed{0\,|\,1\,|\,15\,|\,0}$.

A second contingency table could be formed from the pattern where the 0s occur diagonally in cells 1 and 4 by flip flopping the 1 and the 15 to form $\boxed{0}\ \boxed{15}\ \boxed{1}\ \boxed{0}$.

But each one of these 12 contingency tables reflecting the 15+1+0+0 occupancy pattern can be filled in 16 different ways when we take note of the fact that kangaroos are distinguishable. Remember, we gave each distinguishable kangaroo its own name.

Consider the previous example of the future frequency counts recorded in the contingency table $\boxed{0}\ \boxed{1}\ \boxed{15}\ \boxed{0}$. Alex could be in cell 2 and the other 15 kangaroos in cell 3. Or, Beth could be in cell 2 and the other 15 kangaroos in cell 3. Or, Carl could be in cell 2 and the other 15 kangaroos in cell 3. Or, ... I think you get the picture. There are 16 different ways to fill each of the 12 contingency tables when we take the distinguishability of the kangaroos into account.

So we see that in the end there are $12 \times 16 = 192$ elementary points contributed to the sample space for this particular integer partition. Grinding things down to the finest micro–statement level, *one* elementary point in the sample space might be described by the statement,

Cell 1 "no kangaroos,

Cell 2 Alex,

Cell 3 Beth through Patricia,

Cell 4 no kangaroos."

By applying the formula in Equation (15.2), we implement the above extended verbal explanation for any case desired,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1!\, M_2!\, \cdots\, M_n!} \;=\; \frac{4!}{2!\, 1!\, 0! \cdots 0!\, 1!\, 0!} \times \frac{16!}{15!\, 1!\, 0!\, 0!}$$

$$=\; 12 \times 16$$

$$=\; 192$$

The first term in Equation (15.2) tells us the number of different ways that a particular integer partition can take place. The second term is the multiplicity factor $W(M)$ for the given macro–statement $F_j \equiv (M_1, M_2, \cdots, M_n)$. With 192 elementary points accounted for through the $15 + 1 + 0 + 0$ occupancy pattern, we only have to find the remaining 4,294,967,104 points!

Exercise 15.7.2. Warm up by working on a dice occupancy problem.

Solution to Exercise 15.7.2

As a prelude to compiling the exhaustive listing involving the kangaroos, warm up by practicing on a simpler problem. Suppose the IP wants to predict the first four rolls of an ordinary die. The state space now has dimension $n = 6$ with the sample space determined by setting $M = 4$ for the next four rolls. The total number of elementary points in the sample space is then,

$$n^M = 6^4 = 1296$$

Since the state space consists of six statements, every contingency table consists of six cells. The cells represents the ONE, TWO, $\cdots$, or SIX spots that might appear on a roll of the die.

How many different contingency tables are there for this problem? Equation (15.1) tells us that there are 126 different contingency tables in the die rolling scenario,

$$\begin{aligned}
\frac{(M + n - 1)!}{M!\,(n - 1)!} &= \frac{(4 + 6 - 1)!}{4!\,(6 - 1)!} \\[2mm]
&= \frac{9 \times 8 \times 7 \times 6}{24} \\[2mm]
&= 126
\end{aligned}$$

As we have seen before, the total sum of $M = 4$ rolls can be broken down into five occupancy patterns over the six cells. Each of these five occupancy patterns can be formed in a number of ways. This number of ways that a contingency table could fit one of these occupancy patterns is calculated using the first term,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!}$$

appearing in Equation (15.2).

For example, take the occupancy pattern $4 + 0 + 0 + 0 + 0 + 0$. This pattern indicates that the same face appeared on all four rolls of the die. In other words, getting a THREE on all four rolls fits this pattern. This contingency table would look like $\boxed{0\,|\,0\,|\,4\,|\,0\,|\,0\,|\,0}$.

Obviously, we can imagine six such contingency tables. The formula thankfully also calculates 6 possible contingency tables to reflect the result of obtaining the same face on all four rolls,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times r_t! \times r_M!} = \frac{6!}{5!\,0!\,0!\,0!\,1!} = 6$$

The n in the numerator is the dimension of the state space. In the denominator, the $r_z, r_s, \cdots, r_M$ are the number of cells in which a *zero* count appears, the number of cells in which a *single* count appears, and so on. Notice that the subscript on the final term is M, so that $r_M = 1$, the number of cells in which a frequency count of "4" appears.

Let's do this again. We just examined the first occupancy pattern of $4 + 0 + 0 + 0 + 0 + 0$. The second occupancy pattern that sums to 4 is $3 + 1 + 0 + 0 + 0 + 0$. This stands for obtaining the same face on three rolls, and a different face on the remaining roll. Rolling a FIVE on three rolls and a TWO on the remaining roll would fit this pattern. The contingency table would look like $\boxed{0\,|\,1\,|\,0\,|\,0\,|\,3\,|\,0}$.

The formula tells us that there are now 30 different contingency tables fitting the bill,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times r_t! \times r_M!} = \frac{6!}{4!\ 1!\ 0!\ 1!\ 0!} = 30$$

If we were to carry out the same kind of calculation for the other three occupancy patterns, which are $2 + 2 + 0 + 0 + 0 + 0$, $2 + 1 + 1 + 0 + 0 + 0$, and $1 + 1 + 1 + 1 + 0 + 0$, we would find that there are respectively 15, 60, and 15 different contingency tables associated with each pattern. If we add up all the different ways to form the contingency tables, they should add up to the overall total of 126 contingency tables we calculated at the very beginning,

$$\frac{(M + n - 1)!}{M!\ (n - 1)!} = 6 + 30 + 15 + 60 + 15 = 126$$

When we talked about the kangaroos, the second term in Equation (15.2), the multiplicity factor, told us the number of different ways each contingency table could be formed by taking into account the individual nature of the kangaroos.

When talking about distinguishable dice, they might be four dice of a different color rolled at the same time. Or, we might roll the same die four times and take note on which roll a particular face showed. For the dice rolling scenario, the generic labeling **a, b, c, d** refers not to the names given to kangaroos, but rather to the particular color of the die, say, blue, red, green, or yellow, or to the temporal order of the roll, be it the first, second, third, or fourth roll.

The multiplicity factor $W(M)$ informs us as to the number of ways a particular contingency table might turn out when taking account of the particular color, or the particular temporal order, of the die. For example, we just found out that there were 30 different contingency tables that fit the $3 + 1 + 0 + 0 + 0 + 0$ occupancy pattern. One of those looked like $\boxed{0\,|\,1\,|\,0\,|\,0\,|\,3\,|\,0}$ and represented rolling a TWO and three FIVEs. There are four different ways taking into account the color, or the temporal order, that this particular contingency table could have come about,

$$W(M) = \frac{M!}{M_1!\ M_2!\ M_3!\ M_4!\ M_5!\ M_6!} = \frac{4!}{0!\ 1!\ 0!\ 0!\ 3!\ 0!} = 4$$

Using the labeling system just described, these four situations result in these four elementary points in the sample space,

1. $\boxed{\star}\,\boxed{\text{a}}\,\boxed{\star}\,\boxed{\star}\,\boxed{\text{bcd}}\,\boxed{\star}$

2. $\boxed{\star}\,\boxed{\text{b}}\,\boxed{\star}\,\boxed{\star}\,\boxed{\text{acd}}\,\boxed{\star}$

3. $\boxed{\star}\,\boxed{\text{c}}\,\boxed{\star}\,\boxed{\star}\,\boxed{\text{abd}}\,\boxed{\star}$

4. $\boxed{\star}\,\boxed{\text{d}}\,\boxed{\star}\,\boxed{\star}\,\boxed{\text{abc}}\,\boxed{\star}$

For example, case 2 above could describe the situation of differently colored dice where the blue dice showed a TWO and the red, green, and yellow dice showed a FIVE. Or, it might describe the situation of a different temporal order where TWO appeared on the second roll, and FIVE on the first, third, and fourth rolls.

When we add up all the elementary points in this way we find that,

Elementary points in dice problem =

$$(6 \times 1) + (30 \times 4) + (15 \times 6) + (60 \times 12) + (15 \times 24) = n^M = 6^4 = 1296$$

See Table 15.8 below for a summary of the combinatorial formula of Equation (15.2), and how it accounts for all of the elementary points in the sample space for the dice scenario.

Table 15.8: *Summary table of the five occupancy patterns in the dice rolling scenario. The counts add up to the total number of contingency tables and the total number of elementary points.*

Occupancy Patterns	Multiplicity Factor	Contingency Tables	Elementary Points
$4+0+0+0+0+0$	$\frac{4!}{4!\,0!\,0!\,0!\,0!\,0!} = 1$	$\frac{6!}{5!\,0!\,0!\,0!\,1!} = 6$	$6 \times 1 = 6$
$3+1+0+0+0+0$	$\frac{4!}{3!\,1!\,0!\,0!\,0!\,0!} = 4$	$\frac{6!}{4!\,1!\,0!\,1!\,0!} = 30$	$30 \times 4 = 120$
$2+2+0+0+0+0$	$\frac{4!}{2!\,2!\,0!\,0!\,0!\,0!} = 6$	$\frac{6!}{4!\,0!\,2!\,0!\,0!} = 15$	$15 \times 6 = 90$
$2+1+1+0+0+0$	$\frac{4!}{2!\,1!\,1!\,0!\,0!\,0!} = 12$	$\frac{6!}{3!\,2!\,1!\,0!\,0!} = 60$	$60 \times 12 = 720$
$1+1+1+1+0+0$	$\frac{4!}{1!\,1!\,1!\,1!\,0!\,0!} = 24$	$\frac{6!}{2!\,4!\,0!\,0!\,0!} = 15$	$15 \times 24 = 360$
Sums		126	1296

Exercise 15.7.3. What are some probabilities of dice rolling events under the fair model?

Solution to Exercise 15.7.3

Having gone to all this trouble to provide a detailed listing of every single elementary point, it is fun to pick out various events as aggregates to see what their probabilities would be under the specific model that assigns a numerical value of $Q_i = 1/n$.

Up to now, everything has been simply counting up ways things could happen. There was no mention whatsoever of any probability. However, if we were to assume that every elementary point has the same probability, then the IP might assess the probability for some event as the aggregation of all the elementary points meeting the definition of that event.

For example, what about the event that all four rolls show the same face? The probability for this event would be 6/1296. Or, the event that the same four faces did NOT show up would be assessed as 1290/1296. On this basis, the event that the die showed either all different faces, or two of the same face and the other two rolls a different face has the bulk of the probability at 1080/1296. Furthermore, and somewhat surprisingly, the probability that all different faces appear (360/1296) is less than one–half the probability that, out of the four rolls, two of the faces are the same (810/1296).

Exercise 15.7.4. Begin the full accounting of the sample space in the kangaroo problem.

Solution to Exercise 15.7.4

If you are interested, the next four exercises provide a detailed and exhaustive listing of the elementary points involved in the kangaroo scenario. It follows, albeit with more attention and labor, exactly what was done with the dice.

The entire listing has to be broken down into four tables. Table 15.9 starts off on the next page by listing the occupancy patterns from 16–0–0–0 through 11–2–2–1. Table 15.10 follows with the occupancy patterns from 10–6–0–0 through 9–3–2–2. Table 15.11 is the third breakdown covering the occupancy patterns from 8–8–0–0 through 7–3–3–3. Table 15.12 finishes up with the occupancy patterns from 6–6–4–0 through 4–4–4–4. The notation is changed simply for spacing concerns in the presentation of the tables.

The occupancy patterns appearing in the first column of the following tables represent the partition of an integer, in this case, the partition of 16 into four parts. The number of different contingency tables in the third column is calculated according to the first term in Equation (15.2),

$$\frac{n!}{r_z! \times r_s! \times \cdots \times r_M!}$$

Table 15.9: *The first part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario.*

Occupancy Patterns	Multiplicity Factor		Contingency Tables		Elementary Points		
16–0–0–0	$\frac{16!}{16!\,0!\,0!\,0!} =$	1	$\frac{4!}{3!\,1!} =$ 4		$4 \times$	$1 =$	4
15–1–0–0	$\frac{16!}{15!\,1!\,0!\,0!} =$	16	$\frac{4!}{2!\,1!\,1!} = 12$		$12 \times$	$16 =$	192
14–2–0–0	$\frac{16!}{14!\,2!\,0!\,0!} =$	120	$\frac{4!}{2!\,1!\,1!} = 12$		$12 \times$	$120 =$	1440
14–1–1–0	$\frac{16!}{14!\,1!\,1!\,0!} =$	240	$\frac{4!}{1!\,2!\,1!} = 12$		$12 \times$	$240 =$	2880
13–3–0–0	$\frac{16!}{13!\,3!\,0!\,0!} =$	560	$\frac{4!}{2!\,1!\,1!} = 12$		$12 \times$	$560 =$	6720
13–2–1–0	$\frac{16!}{13!\,2!\,1!\,0!} =$	1680	$\frac{4!}{1!\,1!\,1!\,1!} = 24$		$24 \times$	$1680 =$	40320
13–1–1–1	$\frac{16!}{13!\,1!\,1!\,1!} =$	3360	$\frac{4!}{3!\,1!} =$ 4		$4 \times$	$3360 =$	13440
12–4–0–0	$\frac{16!}{12!\,4!\,0!\,1!} =$	1820	$\frac{4!}{2!\,1!\,1} = 12$		$12 \times$	$1820 =$	21840
12–3–1–0	$\frac{16!}{12!\,3!\,1!\,0!} =$	7280	$\frac{4!}{1!\,1!\,1!\,1!} = 24$		$24 \times$	$7280 =$	174720
12–2–2–0	$\frac{16!}{12!\,2!\,2!\,0!} =$	10920	$\frac{4!}{1!\,2!\,1!} = 12$		$12 \times$	$10920 =$	131040
12–2–1–1	$\frac{16!}{12!\,2!\,1!\,1!} =$	21840	$\frac{4!}{2!\,1!\,1!} = 12$		$12 \times$	$21840 =$	262080
11–5–0–0	$\frac{16!}{11!\,5!\,0!\,0!} =$	4368	$\frac{4!}{2!\,1!\,1} = 12$		$12 \times$	$4368 =$	52416
11–4–1–0	$\frac{16!}{11!\,4!\,1!\,0!} =$	21840	$\frac{4!}{1!\,1!\,1!\,1!} = 24$		$24 \times$	$21840 =$	524160
11–3–2–0	$\frac{16!}{11!\,3!\,2!\,0!} =$	43680	$\frac{4!}{1!\,1!\,1!\,1!} = 24$		$24 \times$	$43680 =$	1048320
11–3–1–1	$\frac{16!}{11!\,3!\,1!\,1!} =$	87360	$\frac{4!}{2!\,1!\,1!} = 12$		$12 \times$	$87360 =$	1048320
11–2–2–1	$\frac{16!}{11!\,2!\,2!\,1!} = 131040$		$\frac{4!}{2!\,1!\,1!} = 12$		$12 \times$	$131040 =$	1572480
Partial Sums			224				4, 900, 372

Since n is small, there is a limited spread in the possible number of contingency tables. The possibilities for the number of contingency tables are, in fact, just 1, 4, 6, 12, or 24. Instead of explicitly showing all seventeen $r_i!$ in the denominator, most of which will be $0! = 1$, only those $r_i!$ not equal to $0!$ will be shown.

For example, examine the very first occupancy pattern of 16–0–0–0, reflecting the fact that all sixteen kangaroos have the same trait. The four possible contingency tables for this are $\boxed{16 \mid 0 \mid 0 \mid 0}$, $\boxed{0 \mid 16 \mid 0 \mid 0}$, $\boxed{0 \mid 0 \mid 16 \mid 0}$, and $\boxed{0 \mid 0 \mid 0 \mid 16}$. Rather than showing,

$$\frac{4!}{3!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 1!} = 4$$

just the condensed form is shown,

$$\frac{4!}{3!\ 1!} = 4$$

In the overall grand accounting, we are attempting to recover the following two sums. The sum over all of the contingency tables must equal 969. The sum over the elementary points must equal 4,294,967,296.

Exercise 15.7.5. Construct the second, third, and final tables listing all of the elementary points in the kangaroo sample space.

Solution to Exercise 15.7.5

Exercise 15.7.4 took on the task of presenting the first table covering the occupancy patterns from $16 + 0 + 0 + 0$ through $11 + 2 + 2 + 1$. Table 15.10 on the next page shows the second part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario. It covers the occupancy patterns from $10 + 6 + 0 + 0$ through $9 + 3 + 2 + 2$.

Table 15.11 shows the third part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario. It covers the occupancy patterns $8 + 8 + 0 + 0$ through $7 + 3 + 3 + 3$. Table 15.12 shows the fourth and final part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario. It covers the occupancy patterns $6 + 6 + 4 + 0$ through $4 + 4 + 4 + 4$.

Adding up the partial sums for the number of contingency tables over the four listings provides the check on the total number of contingency tables,

$$224 + 256 + 340 + 149 = 969$$

Adding up the partial sums for the number of elementary points over the four listings is more onerous, but they do add up to the required,

$$n^M = 4^{16} = 4,294,967,296$$

Table 15.10: *The second part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario.*

Occupancy Patterns	Multiplicity Factor	Contingency Tables	Elementary Points	
10–6–0–0	$\frac{16!}{10!\,6!\,0!\,0!} = 8008$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 8008 =$	96096
10–5–1–0	$\frac{16!}{10!\,5!\,1!\,0!} = 48048$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 48048 =$	1153152
10–4–2–0	$\frac{16!}{10!\,4!\,2!\,0!} = 120120$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 120120 =$	2882880
10–4–1–1	$\frac{16!}{10!\,4!\,1!\,1!} = 240240$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 2420240 =$	2882880
10–3–3–0	$\frac{16!}{10!\,3!\,3!\,0!} = 160160$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 160160 =$	1921920
10–3–2–1	$\frac{16!}{10!\,3!\,2!\,1!} = 480480$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 480480 =$	11531520
10–2–2–2	$\frac{16!}{10!\,2!\,2!\,2!} = 720720$	$\frac{4!}{3!\,1!} = 4$	$4 \times 720720 =$	2882880
9–7–0–0	$\frac{16!}{9!\,7!\,0!\,0!} = 11440$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 11440 =$	137280
9–6–1–0	$\frac{16!}{9!\,6!\,1!\,0!} = 80080$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 80080 =$	1921920
9–5–2–0	$\frac{16!}{9!\,5!\,2!\,0!} = 240240$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 240240 =$	5765760
9–5–1–1	$\frac{16!}{9!\,5!\,1!\,1!} = 480480$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 480480 =$	5765760
9–4–3–0	$\frac{16!}{9!\,4!\,3!\,0!} = 400400$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 400400 =$	9609600
9–4–2–1	$\frac{16!}{9!\,4!\,2!\,1!} = 1201200$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 1201200 =$	28828800
9–3–3–1	$\frac{16!}{9!\,3!\,3!\,1!} = 1601600$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 1601600 =$	19219200
9–3–2–2	$\frac{16!}{9!\,3!\,2!\,2!} = 2402400$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 2402400 =$	28828800
Partial Sums		256	$123,428,448$	

Table 15.11: *The third part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario.*

Occupancy Patterns	Multiplicity Factor	Contingency Tables	Elementary Points
8–8–0–0	$\frac{16!}{8!\,8!\,0!\,0!} = \quad 12870$	$\frac{4!}{2!\,2!} = \quad 6$	$6 \times 12879 = \qquad 77220$
8–7–1–0	$\frac{16!}{8!\,7!\,1!\,0!} = \quad 102960$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 102960 = \qquad 2471040$
8–6–2–0	$\frac{16!}{8!\,6!\,2!\,0!} = \quad 360360$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 360360 = \qquad 8648640$
8–6–1–1	$\frac{16!}{8!\,6!\,1!\,1!} = \quad 720720$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 720720 = \qquad 8648640$
8–5–3–0	$\frac{16!}{8!\,5!\,3!\,0!} = \quad 720720$	$\frac{4!}{2!\,1!\,1!\,1!} = 24$	$24 \times 720720 = \qquad 17297280$
8–5–2–1	$\frac{16!}{8!\,5!\,2!\,1!} = \quad 2162160$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 2162160 = \qquad 51891840$
8–4–4–0	$\frac{16!}{8!\,4!\,4!\,0!} = \quad 900900$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 900900 = \qquad 10810800$
8–4–3–1	$\frac{16!}{8!\,4!\,3!\,1!} = \quad 3603600$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 3603600 = \qquad 86486400$
8–4–2–2	$\frac{16!}{8!\,4!\,2!\,2!} = \quad 5405400$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 5405400 = \qquad 64864800$
8–3–3–2	$\frac{16!}{8!\,3!\,3!\,2!} = \quad 7207200$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 7207200 = \qquad 86486400$
7–7–2–0	$\frac{16!}{7!\,7!\,2!\,0!} = \quad 411840$	$\frac{4!}{2!\,1!} = 12$	$12 \times 411840 = \qquad 4942080$
7–7–1–1	$\frac{16!}{7!\,7!\,1!\,1!} = \quad 823680$	$\frac{4!}{2!\,2!} = \quad 6$	$6 \times 823680 = \qquad 4942080$
7–6–3–0	$\frac{16!}{7!\,6!\,3!\,0!} = \quad 960960$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 960960 = \qquad 23063040$
7–6–2–1	$\frac{16!}{7!\,6!\,2!\,1!} = \quad 2882880$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 2882880 = \qquad 69189120$
7–5–4–0	$\frac{16!}{7!\,5!\,4!\,0!} = \quad 1441440$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 1441440 = \qquad 34594560$
7–5–3–1	$\frac{16!}{7!\,5!\,3!\,1!} = \quad 5765760$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 5765760 = \qquad 138378240$
7–5–2–2	$\frac{16!}{7!\,5!\,2!\,2!} = \quad 8648640$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 8648640 = \qquad 103783680$
7–4–4–1	$\frac{16!}{7!\,4!\,4!\,1!} = \quad 7207200$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 7207200 = \qquad 86486400$
7–4–3–2	$\frac{16!}{7!\,4!\,3!\,2!} = 14414400$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 14414400 = \quad 345945600$
7–3–3–3	$\frac{16!}{7!\,3!\,3!\,3!} = 19219200$	$\frac{4!}{3!\,1!} = 4$	$4 \times 19219200 = \qquad 76876800$
Partial Sums		340	$1,225,884,660$

Table 15.12: *The fourth part of the exhaustive listing of the elementary points in the sample space for the kangaroo scenario.*

Occupancy Patterns	Multiplicity Factor	Contingency Tables	Elementary Points
6–6–4–0	$\frac{16!}{6!\,6!\,4!\,0!} = 1681680$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 1681680 = 20180160$
6–6–3–1	$\frac{16!}{6!\,6!\,3!\,1!} = 6726720$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 6726720 = 80720640$
6–6–2–2	$\frac{16!}{6!\,6!\,2!\,2!} = 10090080$	$\frac{4!}{2!\,2!} = 6$	$6 \times 10090080 = 60540480$
6–5–5–0	$\frac{16!}{6!\,5!\,5!\,0!} = 2018016$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 2018016 = 24216192$
6–5–4–1	$\frac{16!}{6!\,5!\,4!\,1!} = 10090080$	$\frac{4!}{2!\,1!\,1!\,1!} = 24$	$24 \times 10090080 = 242161920$
6–5–3–2	$\frac{16!}{6!\,5!\,3!\,2!} = 20180160$	$\frac{4!}{1!\,1!\,1!\,1!} = 24$	$24 \times 20180160 = 484323840$
6–4–4–2	$\frac{16!}{6!\,4!\,4!\,2!} = 25225200$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 25225200 = 302702400$
6–4–3–3	$\frac{16!}{6!\,4!\,3!\,3!} = 33633600$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 33633600 = 403603200$
5–5–5–1	$\frac{16!}{5!\,5!\,5!\,1!} = 12108096$	$\frac{4!}{3!\,1!} = 4$	$4 \times 12108096 = 48432384$
5–5–4–2	$\frac{16!}{5!\,5!\,4!\,2!} = 30270240$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 30270240 = 363242880$
5–5–3–3	$\frac{16!}{5!\,5!\,3!\,3!} = 40360320$	$\frac{4!}{2!\,2!} = 6$	$6 \times 40360320 = 242161920$
5–4–4–3	$\frac{16!}{5!\,4!\,4!\,3!} = 50450400$	$\frac{4!}{2!\,1!\,1!} = 12$	$12 \times 50450400 = 605404800$
4–4–4–4	$\frac{16!}{4!\,4!\,4!\,4!} = 63063000$	$\frac{4!}{4!} = 1$	$1 \times 63063000 = 63063000$
Partial Sums		149	$2,940,753,816$

For fun, use these counts of the number of elementary points in the sample space to stipulate any probability under the fair model. What is the probability of obtaining eight kangaroos who share the same beer–hand preference without being concerned about the preferences of the other eight kangaroos? Refer back to Table 15.11 to add up all the relevant elementary points. The probability is,

$$\frac{337,683,060}{4,294,967,296} = 0.0786$$

One example of the 24 contingency tables available from the occupancy pattern $8 + 5 + 2 + 1$ is the contingency table $\boxed{2\;|\;5\;|\;8\;|\;1}$. Here, eight kangaroos share the same preference for drinking Foster's with the left hand. One possibility that might have occurred, now taking note of the distinguishable kangaroos by listing their names, is that the eight kangaroos who preferred to drink Foster's with their left hand were Beth, Edgar, George, Jerri, Klaus, Mitchell, Nicki, and Patricia.

Exercise 15.7.6. Trace an elementary sample point back up the chain as a member of higher and higher sets of aggregated sample points.

Solution to Exercise 15.7.6

For a final exercise in the same spirit, move up in the hierarchy from the finest grain to the coarsest grain. In other words, trace the path beginning at a particular micro–statement, that is, from one elementary point in the sample space, all the way back up to some final event consisting of this elementary sampling point together with countless other elementary sampling points just like it according to the criterion defining the event.

For example, how many elementary sampling points comprise the event that no more than five kangaroos share the same trait?

Suppose, for the sake of the example, that we begin with the following micro–statement concerning what we might observe about the kangaroos and their hand–beer preference. This elementary sampling point is illustrated in Table 15.13.

This is as detailed a statement as we can make about the sample space given the state space and the number of future observations. We have identified by name each single kangaroo and its particular hand–beer preference.

The contingency table for these observations on the sixteen kangaroos is then $\boxed{3\;|\;4\;|\;5\;|\;4}$. But this particular set of frequency counts is only one example of twelve such contingency tables following the $5 + 4 + 4 + 3$ occupancy pattern. Refer back to Table 15.12 where this occupancy pattern occurs in the next to the last row. So we have moved up to considering the situation where five kangaroos share the same trait, four share a second trait, and so on.

If the IP were to voluntarily discard the information about which particular kangaroos possessed the traits, this $\boxed{3\;|\;4\;|\;5\;|\;4}$ contingency table could have arisen

Table 15.13: *A micro–statement that details the hand–beer preference for sixteen individual kangaroos. The cell of the contingency table in which they would be placed is shown in the final column.*

Kangaroo	Name	Preference	Cell
1	Alex	$\overline{R}\,\overline{F}$	4
2	Beth	$\overline{R}F$	3
3	Carl	$\overline{R}\,\overline{F}$	4
4	Dawn	RF	1
5	Edgar	$\overline{R}F$	3
6	Florence	$R\overline{F}$	2
7	George	$\overline{R}F$	3
8	Helen	$R\overline{F}$	2
9	Igor	$\overline{R}\,\overline{F}$	4
10	Jerri	$\overline{R}F$	3
11	Klaus	RF	1
12	Leonora	$\overline{R}\,\overline{F}$	4
13	Mitchell	$\overline{R}F$	3
14	Nicki	$R\overline{F}$	2
15	Oscar	RF	1
16	Patricia	$R\overline{F}$	2

in $50, 450, 400$ different ways. Moving up yet one more level in the hierarchy, there are $12 \times 50, 450, 400 = 605, 404, 800$ elementary points in the sample space that fit the description of five kangaroos with one hand–beer preference, four kangaroos with a second hand–beer preference, four kangaroos possessing the third trait, and three kangaroos possessing the final trait.

We can now move up to the highest level of aggregated sample points and ask about the number of elementary points with no more than five kangaroos as a frequency count in any one cell. Now there are, as shown at the top of the next page in Table 15.14, $1, 322, 304, 984$ sample points from the total of 4,294,967,296.

Under the "fair" model where there are no correlations among the traits, and a numerical value of $Q_i = 1/4$ is assigned to the probability that a kangaroo has a particular beer–hand preference, this event accounts for over 30% of the total probability for any conceivable set of frequency counts.

You might have expected to see such generally spread out frequency counts under such a fair model with no correlations. But this does not imply that the "maximum entropy frequency count" (wrong!) reflected in the contingency table $\boxed{4}\,\boxed{4}\,\boxed{4}\,\boxed{4}$ is particularly more probable than other high–level events. A quick glance at Table 15.14 confirms this.

Table 15.14: *Counting up the number of elementary points comprising the situation where no more than five kangaroos are placed in any cell of the contingency table.*

Occupancy pattern	Sample points
$5 + 5 + 5 + 1$	48,432,384
$5 + 5 + 4 + 2$	363,242,880
$5 + 5 + 3 + 3$	242,161,920
$5 + 4 + 4 + 3$	605,404,800
$4 + 4 + 4 + 4$	63,063,000
Sum	1,322,304,984

Exercise 15.7.7. What pattern of seven accidents during a week is more indicative of the fair model ascribing equal probability to the occurrence of an accident?

Solution to Exercise 15.7.7

See Feller's [4] solved example of this combinatorial problem on his pages 39–40. I employed exactly the same technique for the kangaroos in this Chapter except that I chose to present the computations in a more direct and transparent manner than Feller.

If the probability of an accident on any given day is $1/7$, then it is far more likely that one will observe weeks with two or three accident–free days as opposed to an accident every day.

Exercise 15.7.8. As a complement to the numerical exercises, show, in general, how the probabilities for future frequency counts may approach interesting values.

Solution to Exercise 15.7.8

As opposed to my numerical explorations, Jaynes actually constructed symbolic solutions, with somewhat less impact in my opinion. Consider, for example, the situation where we want to calculate a state of knowledge for the future frequency count where all sixteen kangaroos are observed to be right–handed Corona drinkers. In other words, all $M = 16$ kangaroos will be placed into cell 2 of the contingency table with frequency counts in each of the four cells,

$$M_1 = 0, M_2 = 16, M_3 = 0, M_4 = 0$$

This particular situation was treated in the numerical exercise in section 15.4.5 and Table 15.5. There we saw that as $\alpha_2 \to \infty$ while α_1, α_3, and α_4 remained fixed at 1, the probability for the contingency table $\boxed{0\ |\ 16\ |\ 0\ |\ 0}$ approached 1.

In this example and others, as Jaynes pointed out (Equation (37) in MKN), one can see symbolically that the probability of the frequency configuration $\boxed{0\,|\,16\,|\,0\,|\,0}$ must approach 1 under the stated conditions for the α_i. If the probability of this particular frequency count is approaching 1 then, of course, all the other contingency tables combined must be approaching 0.

Refer back to Equation (15.3). For our current numerical example, the dimension of the state space is $n = 4$, so the equation will consist of the constant term and four other terms. The first, third, and fourth terms where the $\alpha_i = 1$ are all equal to 1 because,

$$\frac{(M_i + 1 - 1)!}{M_i!\, 0!} = 1$$

where $i = 1, 3, 4$. We only have to be concerned with the constant term and the second term where $\alpha_2 \to \infty$,

$$P(M_1, M_2, M_3, M_4) = \frac{M!\,(\mathcal{A} - 1)!}{(M + \mathcal{A} - 1)!} \times \frac{(M_2 + \alpha_2 - 1)!}{M_2!\,(\alpha_2 - 1)!}$$

Since $M_2 = M$ for this frequency count of interest,

$$
\begin{aligned}
P(M_1 = 0, M_2 = M, M_3 = 0, M_4 = 0) &= \frac{(\mathcal{A} - 1)!}{(M + \mathcal{A} - 1)!} \times \frac{(M + \alpha_2 - 1)!}{(\alpha_2 - 1)!} \\[2mm]
&= \frac{\Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \frac{\Gamma(M + \alpha_2)}{\Gamma(\alpha_2)}
\end{aligned}
$$

Since $\alpha_2 \to \infty$, α_2 also $\to \mathcal{A}$ and therefore,

$$
\begin{aligned}
P(M_1 = 0, M_2 = M, M_3 = 0, M_4 = 0) &\to \frac{\Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \frac{\Gamma(M + \mathcal{A})}{\Gamma(\mathcal{A})} \\[2mm]
&\to 1
\end{aligned}
$$

Exercise 15.7.9. Gather together in one place a detailed derivation of Equation (15.3).

Solution to Exercise 15.7.9

It is best, I believe, to begin with the conceptual template of the marginalization over a discrete number of models so that the integration symbology does not become too distracting right at the very start.

If the probability for a statement A is going to be thought about through the intervention of parametric models, then in probability theory the parameters always get swept under the rug through marginalization. So, simply and generically, this is the conceptual template to keep in mind for the derivation to follow.

$$P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k) \tag{15.4}$$

$(A = a_i)$ is just a statement about a kangaroo possessing one of the four available beer–hand preference traits. In other words, it is one of the n joint statements in the state space.

When devising macro–statements involving frequency counts of the kangaroos appearing in the various cells of the contingency table, this basic statement about one kangaroo gets repeated many times. Furthermore, the summation over a finite number of models is replaced by a multiple integration over the entire model space. This is indicated by showing the region of integration, as an index to the integral sign, by $\sum q_i = 1$.

Each q_i will be changed by an infinitesimal amount during the integration, but the sum of the q_i must remain equal to 1. When, during the course of the multiple integration, q_1 eventually takes on, say, the value 0.6, then q_2 is restricted to taking on values in the range $0 < q_2 \leq 0.4$ with subsequent restrictions on q_3 and q_4. Furthermore, only $n - 1$ values of the q_i need be specified since the nth will then be determined. Aesthetically though, it is more pleasing to show all n q_i in the equations.

The proof is no harder if it is written in terms of general n. The final formula for any specific dimension of the state space can be substituted into the general formula at the end.

Taking all of this into account, the simple conceptual template for the prior predictive formula shown as Equation (15.4) gets turned into,

$$P(M_1, M_2, \cdots, M_n) = \int \cdots \int_{\sum q_i = 1} P(M_1, M_2, \cdots, M_n \,|\, \mathcal{M}) \, P(\mathcal{M}) \, \mathrm{d}\mathcal{M} \quad (15.5)$$

The models $\mathcal{M}$ are statements asserting that the "correct" numerical values were assigned to the probabilities for the n joint statements comprising the state space. Then Equation (15.5) is less sloppily written as,

$$P(M_1, M_2, \cdots, M_n) =$$

$$\int \cdots \int_{\sum q_i = 1} P(M_1, M_2, \cdots, M_n \,|\, q_1, q_2, \cdots, q_n) \, \mathrm{pdf} \,(q_1, q_2, \cdots, q_n) \, \mathrm{d}q_i \quad (15.6)$$

It is very important here to emphasize that in transitioning from Equation (15.5) to Equation (15.6), we have not included an expression involving a "probability of a probability." Such a concept is anathema to any theory of probability.

The expression $\mathrm{pdf}\,(q_1, q_2, \cdots, q_n)$ is legitimate because it is just another way of expressing the probability for a *statement* $\mathcal{M}_k$. The statement might be phrased in this colloquial manner. "I am model $\mathcal{M}_k$ and I am making the following correct numerical assignments $q_1, q_2, \cdots, q_n$ to the probabilities for the n joint statements in the state space." $P(\mathcal{M}_k)$ is then the IP's degree of belief that such a statement about the assignments is TRUE.

Continuing on with the derivation, the probability for the future frequency counts under the assumption that some model has made legitimate numerical assignments to the probabilities in the state space is the multinomial distribution,

$$P(M_1, M_2, \cdots, M_n \,|\, \mathcal{M}_k) = \frac{M!}{M_1! \; M_2! \; \cdots \; M_n!} \times q_1^{M_1} \, q_2^{M_2} \; \cdots \; q_n^{M_n}$$

where, in the upcoming equations, the multiplicity factor has the notation,

$$W(M) \equiv \frac{M!}{M_1! \; M_2! \; \cdots \; M_n!}$$

After making this substitution, the first term is then written as,

$$P(M_1, M_2, \cdots, M_n \,|\, \mathcal{M}_k) = W(M) \times q_1^{M_1} \, q_2^{M_2} \; \cdots \; q_n^{M_n}$$

The probability density function for the models is given by the Dirichlet distribution below with parameters indicated by α_i.

$$\mathrm{pdf}\,(q_1, q_2, \cdots, q_n) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \, q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1}$$

After making this substitution for $\mathrm{pdf}\,(q_1, q_2, \cdots, q_n)$, the derivation continues on as,

$$P(M_1, M_2, \cdots, M_n) =$$

$$\int \cdots \int_{\sum q_i = 1} W(M) \, q_1^{M_1} \, q_2^{M_2} \; \cdots \; q_n^{M_n} \, \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \, q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1} \, \mathrm{d}q_i$$

$$= \; W(M) \int \cdots \int_{\sum q_i = 1} q_1^{M_1} \, q_2^{M_2} \; \cdots \; q_n^{M_n} \, \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \, q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1} \, \mathrm{d}q_i$$

$$= \; W(M) \, \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \int \cdots \int_{\sum q_i = 1} q_1^{M_1} \, q_2^{M_2} \; \cdots \; q_n^{M_n} \, q_1^{\alpha_1 - 1} \, q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1} \, \mathrm{d}q_i$$

$$= \; W(M) \, \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \int \cdots \int_{\sum q_i = 1} q_1^{M_1 + \alpha_1 - 1} \, q_2^{M_2 + \alpha_2 - 1} \; \cdots \; q_n^{M_n + \alpha_n - 1} \, \mathrm{d}q_i$$

The constant factors not depending on the integration variables q_i have been pulled out from underneath the integral. The integral remaining is a Dirichlet *integral* with the obvious solution (if the form of the Dirichlet *probability distribution* is accepted as correct),

$$\int \cdots \int_{\sum q_i = 1} q_1^{M_1 + \alpha_1 - 1} \, q_2^{M_2 + \alpha_2 - 1} \; \cdots \; q_n^{M_n + \alpha_n - 1} \, \mathrm{d}q_i = \frac{\prod_{i=1}^n \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^n M_i + \alpha_i)}$$

After making this final substitution, we now have,

$$P(M_1, M_2, \cdots, M_n) = W(M) \times \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^n \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^n M_i + \alpha_i)} \tag{15.7}$$

Let us stop here momentarily at Equation (15.7) and see what happens when all the α_i are set equal to 1. This setting of the α_i parameters represents a uniform distribution over the model space. Converting the Gamma functions back into factorials yields this result,

$$P(M_1, M_2, \cdots, M_n) \;\; = \;\; \frac{M!}{M_1! \; M_2! \; \cdots \; M_n!} \times \frac{(n-1)!}{0! \cdots 0!} \times$$

$$\frac{M_1! \; M_2! \; \cdots \; M_n!}{[(M_1 + 1 + M_2 + 1 + \cdots + M_n + 1) - 1]!}$$

$$P(M_1, M_2, \cdots, M_n) \;\; = \;\; \frac{M! \, (n-1)!}{(M + n - 1)!} \tag{15.8}$$

When every model has the same relative status, the probability of any future frequency count does not depend on the particular M_i, but is the same for all M_i. In our kangaroo example this got translated into,

$$P(M_1 = 16, M_2 = 0, M_3 = 0, M_4 = 0) = P(M_1 = 4, M_2 = 4, M_3 = 4, M_4 = 4) = 1/969$$

Whether all 16 kangaroos are crammed into the same cell, or spread evenly over all four cells, makes no difference. The probability for either contingency table is the same. And the inverse of the result in Equation (15.8) also tells us the total number of contingency tables. This result appears in Jaynes as his Equation (32).

Our goal is now to complete the derivation as originally begun without the detour looking at the consequences of a uniform distribution over the model space. Consider any general α_i and not just the specific case of all $\alpha_i = 1$. The final result will look like,

$$P(M_1, M_2, \cdots, M_n) = \frac{M! \, (\mathcal{A} - 1)!}{(M + \mathcal{A} - 1)!} \prod_{i=1}^{n} \frac{(M_i + \alpha_i - 1)!}{M_i! \, (\alpha_i - 1)!} \tag{15.9}$$

This is Jaynes's Equation (31).

We pick up the derivation at the spot right before we made the $\alpha_i = 1$ assumption. Keep in mind that the conceptual template of,

$$P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i \,|\, \mathcal{M}_k) \, P(\mathcal{M}_k)$$

is still in effect. Start by adding up whatever the α_i have been assigned and substituting,

$$\sum_{i=1}^{n} \alpha_i = \mathcal{A}$$

$$
\begin{aligned}
P(M_1, M_2, \cdots, M_n) &= W(M) \, \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \, \frac{\prod_{i=1}^{n} \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^{n} M_i + \alpha_i)} \\[2mm]
&= W(M) \, \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \, \frac{\prod_{i=1}^{n} \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^{n} M_i + \alpha_i)}
\end{aligned}
$$

Take the next easy step and substitute the factorial expression for the multiplicity factor $W(M)$,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!}{M_1! \, M_2! \, \cdots M_n!} \, \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \, \frac{\prod_{i=1}^{n} \Gamma(M_i + \alpha_i)}{\Gamma(\sum_{i=1}^{n} M_i + \alpha_i)}$$

Now work on the denominator in the last term to substitute for the summations,

$$
\begin{aligned}
\Gamma \left[\sum_{i=1}^{n} (M_i + \alpha_i) \right] &= \Gamma \left[\sum_{i=1}^{n} M_i + \sum_{i=1}^{n} \alpha_i \right] \\[2mm]
&= \Gamma(M + \mathcal{A})
\end{aligned}
$$

Substitute this into the main expression,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!}{M_1! \, M_2! \, \cdots \, M_n!} \, \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \, \frac{\prod_{i=1}^{n} \Gamma(M_i + \alpha_i)}{\Gamma(M + \mathcal{A})}$$

At this point, we want to explicitly show the expansion of the products in the denominator of the second term and the numerator of the third term.

$$P(M_1, M_2, \cdots, M_n) = \frac{M!}{M_1! \, M_2! \, \cdots \, M_n!} \times \frac{\Gamma(\mathcal{A})}{\Gamma(\alpha_1) \, \Gamma(\alpha_2) \cdots \Gamma(\alpha_n)} \times$$

$$\frac{\Gamma(M_1 + \alpha_1) \, \Gamma(M_2 + \alpha_2) \cdots \Gamma(M_n + \alpha_n)}{\Gamma(M + \mathcal{A})}$$

We are almost finished. The final manipulation that Jaynes does is to rearrange these three terms for later purposes. The first term will contain M and $\mathcal{A}$, while the other n terms will contain M_i and α_i.

$$P(M_1, M_2, \cdots, M_n) =$$

$$\frac{M! \, \Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \frac{\Gamma(M_1 + \alpha_1)}{M_1! \, \Gamma(\alpha_1)} \times \frac{\Gamma(M_2 + \alpha_2)}{M_2! \, \Gamma(\alpha_2)} \times \cdots \times \frac{\Gamma(M_n + \alpha_n)}{M_n! \, \Gamma(\alpha_n)} =$$

$$\frac{M! \, \Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \prod_{i=1}^{n} \frac{\Gamma(M_i + \alpha_i)}{M_i! \, \Gamma(\alpha_i)} \tag{15.10}$$

The only difference between Equation (15.10) and Jaynes's Equation (31) is that he used the notation k_i and K for obvious reasons instead of my more generic α_i and $\mathcal{A}$. In addition, I prefer to employ the notation of M_i to distinguish the future frequency counts from the past frequency counts, N_i.

If we were content to simply look at integer values of α_i, then this equation could be turned into one involving all factorial functions as in Equation (15.9),

$$P(M_1, M_2, \cdots, M_n) = \frac{M! \, (\mathcal{A} - 1)!}{(M + \mathcal{A} - 1)!} \prod_{i=1}^{n} \frac{(M_i + \alpha_i - 1)!}{M_i! \, (\alpha_i - 1)!}$$

But to the computer, calculating the Gamma function is just as easy as the factorial, so we leave the formula in the form of Equation (15.10), (same as Equation (15.3)), to handle non–integer values of α_i. The *Mathematica* program implementing this function is discussed in Appendix E.

Exercise 15.7.10. What are the startling implications if an IP has to make an inference when it is "totally ignorant?"

Solution to Exercise 15.7.10

The enormous disparity in the multiplicity factor for different frequency counts is completely irrelevant when the IP adopts a uniform distribution over model space. When it is desired that every conceivable model for numerical assignments be granted equal status, contingency tables with all 16 kangaroos sharing the same beer–hand preference are just as likely to be true as contingency tables where they are spread out over the available traits.

This phenomenon takes place despite the fact that the contingency table where all the kangaroos are crammed into one cell can happen in only one way, while more spread out arrangements can happen in many millions of ways.

When the IP is laboring under total ignorance, it is indeed uninformed about all of the models. However, the results highlighted in this Chapter just as clearly show that the IP is not equally uninformed about the elementary sample points.

The probability for a future frequency count of $\boxed{16\,|\,0\,|\,0\,|\,0}$ which consists of just *one* elementary sample point is the same at a value of 1/969 as the probability for a future frequency count of $\boxed{4\,|\,4\,|\,4\,|\,4}$ which consists of over 63 million elementary sample points.

Under Haldane's models for total ignorance, things are even more dramatic. The probability for the first case approaches 1/4, while the probability for the second case approaches 0. Once again, this happens despite the enormous disparity in their respective multiplicity factors. Total ignorance about perfect causal linkage models leads to an exclusive concentration on just four sample points.

There really is a marked distinction between an ontological fact concerning the total number of ways things could possibly happen, and an epistemological fact concerning an IP's state of knowledge about how those things might happen.

Chapter 16

Predicting the Behavior of Cellular Automata?

16.1 Introduction

The goal of this first Volume was to gently introduce the probabilistic basis for information processing. We conclude this introduction to inferencing, probability, and information processing with another look at cellular automata.

Specifically, we discuss cellular automata from the viewpoint of probability, and, by doing so, also take the liberty of calling this perspective a generalization of deterministic CA. Unfortunately, at this stage of the game, our conclusions take on a rather pessimistic tone. It is very hard to use the lessons learned so far to predict how CA will behave as they evolve over time.

As was becoming evident to us in some of our last few examples, combinatorial explosion is a real hindrance to further progress. We are going to bump up against this insidious effect in full force as we develop the theme of this final Chapter.

Discussing deterministic CA and probability in the same breath is interesting because of the clash of fundamental concepts. I have previously used the label *probabilistic CA*, but such a term is really a classic case of an oxymoron.

Cellular automata are, buying into Wolfram's arguments, the simplest possible stand–ins for an ontological system. Such systems are, by definition, deterministic, but they may also emulate other ontological systems whose outward appearance may seem quite removed from the simple abstractness of an elementary cellular automaton.

Certainly, such systems must be directly computed to observe their behavior. Thus, these CA have nothing whatsoever to do with inferencing, probability, or information processing. I have been at pains to constantly refer to them as deductive systems.

Probability enters into the discussion about CA when an information processor looks at some CA, or ontological system, and admits that there exists some missing information in *its* knowledge about the system. Thus, probability can only be brought in when there is uncertainty at the epistemological level in an information processor's state of knowledge.

Consider the following thought experiment. An alien IP has constructed some ontological system to follow the rules of a CA. This alien IP has tried to make it very complicated, say, where the many colors of an updated cell depend on a seven argument Boolean function. As far as the alien IP is concerned, this system is a deductive, deterministic system. Hence, it never has to refer to the notions of inferencing, probability, or information processing. For it, there is no missing information.

On the other hand, you and I, also acting as information processors, are looking at the output of this CA. We are overwhelmed by our lack of information about the peculiar behavior of this CA. All we can do is attempt to make inferences using probability and information processing.

Originally, my intent in introducing elementary cellular automata was because they were an interesting example of a Boolean Algebra defined on three arguments and, thus, a very modest extension of Classical Logic. The fact that Wolfram (and his colleagues) could establish that at least one of these elementary CA was a Universal Turing Machine was a remarkable *tour–de–force*. But he reckoned that one of the primary distinguishing characteristics of an elementary CA following Rule 110 is the very fact that one cannot *predict* what effects will ensue far into the future!

Now, to my way of thinking, this is an astonishing property of CA. If we employ something that is purported to be a simple and general computational engine, we are nonetheless prevented from knowing what it will produce after it has been chugging away for some time! The implications seem to be that we can never simulate the Universe faster than it itself is actually computing reality. Therefore, we are forever forbidden from lifting the veil on its future evolution!

This Chapter gathers together the material on predicting future events using the formal rules from probability theory as discussed in the last few Chapters. These rules are then applied to cellular automata because we set up situations where there is missing information.

Earlier, the formal rules for manipulating probabilities were shown to generalize Classical Logic. Furthermore, they were shown to generalize deterministic CA in a straightforward manner.

I do like what Wolfram ([20], pg, 922) had to say about probabilistic cellular automata within the context of a discussion about continuous cellular automata:

> As an alternative to having continuous values at each cell, one can consider ordinary cellular automata with discrete values, but introduce probabilities for, say, two different rules to be applied at each cell.

He is essentially repeating here the very essence of the prediction formula.

Any particular model $\mathcal{M}_k$, (and here those $\mathcal{M}_k$ consist of the 256 rules for elementary CA), determines how the current cell B_{N+1}, will be updated depending on its value, B_N, and the value of its two neighbors, A_N and C_N, at the immediately previous time step. Explicitly, what any model does is assign numerical values to the 16 cells of the joint probability table involving B_{N+1}, B_N, A_N, and C_N. We saw some examples of this back in Chapter Nine.

Previous Chapters have discussed in detail what happens with the prediction formula when dealing with things like flips of a coin, rolls of a die, beer drinking kangaroos, and the graduation success of college students. Causal factors were given more and more prominence with each successive scenario. Now it is time for a try at predicting the future evolution of a CA. The previous cell and its two neighbors are viewed as the causal factors, and some past data are also available.

16.2 Revisiting the Logic Functions

As preparation for the exercises to come, and as a final farewell to the 16 logic functions with which we began this Volume, consider this example. It consists of an even simpler deterministic CA when compared to Wolfram's elementary set of 256 CA.

A cell's color is updated by looking at its own previous value and the color of its one neighbor to the left. Suppose the IP wished to assess its state of knowledge that the updated cell is black given that it was white at the previous time step, and its neighbor to the left was black.

This situation might be characterized with any of the following three probability expressions,

Expression 1. $P(B_{N+1} = b \,|\, A_N = b, B_N = w, \text{Rule } 13)$

Expression 2. $P(Z = T \,|\, A = T, B = F, A \to B)$

Expression 3. $P(A = a_1 \,|\, B = b_1, C = c_2, \mathcal{M}_k)$

The first expression was used for CA, the second for logic functions, and the third for a statement conditioned on two causal factors. We see the presence, as well, of either a CA rule, a logic function, or some k^{th} model.

We'll conduct the exercise in this section in terms of the second expression emphasizing the logic functions. The state space has dimension of $n = 8$. The joint probability table will consist of eight cells to index the probability assigned to each one of the eight possible joint statements involving Z, A, and B.

The two arguments to the logic function are A and B, and each can assume the value of T or F. The functional assignment Z can also assume only the values of T or F. Select the implication logic function, $A \to B$, $f_{13}(A, B)$, which has a

functional assignment of F only if $A = T$ and $B = F$. In the other three cases, the functional assignment is T. All of this is familiar from the earlier Chapters.

There is zero probability, as found from either Bayes's Theorem, or a direct application of the definition of implication, that $Z = T$ given the stated conditions. From the CA perspective, the ontological system will not update the color of the cell to black. Instead, it is certain that system will update the cell to white.

If the IP is completely uninformed about which of the sixteen logic functions is operating to control the output, then it must average the output of all the logic functions as weighted by the probability of the logic function. If the IP is truly "ignorant," then it must assign equal weight to all of the logic functions.

This averaging over the probability for the 16 logic functions looks like,

$$P(Z = T \mid A = T, B = F) =$$

$$\sum_{j=1}^{16} P(Z = T \mid A = T, B = F, f_j\,[A, B]\,) \times P(f_j\,[A, B]\,)$$

Exercise 16.9.1 returns the answer that,

$$P(Z = T \mid A = T, B = F) = 1/2$$

which is eminently reasonable. Because if it is not known which logic function is applicable, then it makes sense that knowledge about the functional output should be maximally non–committal.

If some previous data exist, the IP can leverage it to change the relative status of the importance of the logic functions that control the output. As usual, resort to this formula,

$$P(Z \mid A, \overline{B}, \mathcal{D}) = \sum_{j=1}^{16} P(Z \mid A, \overline{B}, f_j\,[A, B]\,) \times P(f_j\,[A, B] \mid \mathcal{D})$$

Suppose that it were observed that a logic function provided a functional assignment of T when its two arguments were $A = T$ and $B = T$. Now the IP has one data point, $\mathcal{D} = \{Z_1 = T, A_1 = T, B_1 = T\}$.

The IP asks, "What is my state of knowledge, $P(Z_2 = T \mid A_2 = T, B_2 = F, \mathcal{D})$, concerning the truth of the statement that the functional output is T now that I have in my possession this one observation?" Exercise 16.9.2 shows that the state of knowledge doesn't change; it remains at $1/2$. Even though the IP was no longer in a state of complete ignorance about the logic functions, it still could not increase its degree of belief about the output.

16.3   Probabilities and Three Causal Factors

Of course, there is no difficulty in writing out the probability for some statement given *three* causal factors and a model. Different expressions for these probabilities serve to emphasize different viewpoints as shown in the last section. However, there is an obvious conceptual probabilistic similarity underlying diverse phenomena such as Boolean functions, cellular automata, and college graduation.

We first explored the generality of Boolean functions back in Chapter Three. There, we defined the 256 possible Boolean functions with three arguments. We then showed how these Boolean functions matched up with Wolfram's set of 256 elementary cellular automata (ECA).

The probability of an updated cell, the output of a Boolean function, or even a student's graduation might now be written with the following three expressions,

Expression 1. $P(B_{N+1} = b \,|\, A_N = b, B_N = w, C_N = w, \text{Rule 110})$

Expression 2. $P(Z = T \,|\, A = T, B = F, C = F, f_\star[A, B, C])$

Expression 3. $P(A = a_1 \,|\, B = b_1, C = c_2, D = d_2, \mathcal{M}_k)$

During the course of Volume I, we have looked at many examples following these templates. The state space has increased from $n = 8$ to $n = 16$, but when confined to one model, the numerical assignments to such a 16 cell joint probability table were manageable.

We did have to face the unpleasant reality that under the prediction equation's demand that we average over the model space, we would have to fill in these 16 cells many, many times over with different numerical assignments. Hence, our understandable elation over the fact that the integration over the q_i in the prediction formulas took care of all of this in one fell swoop.

Back in Chapter Nine, we broached the problem of averaging over the model space for cellular automata. The following formula was easily written down from the formal rule template for assessing a state of knowledge about the color of a cell in an ECA about to be updated.

$$P(B_{N+1} \,|\, A_N, B_N, C_N) = \sum_{k=1}^{256} P(B_{N+1} \,|\, A_N, B_N, C_N, \text{Rule}_k) \, P(\text{Rule}_k)$$

In a simple illustrative example given there, the probability that the cell will be colored white was $1/2$ when averaged over just two models, Rule 110 and Rule 30.

16.4   Conditional Independence

Now, the most obvious thing we would like to do is generalize in the direction of predicting the next two cells to be updated after some data have been observed.

$$P(B_{N+1}, C_{N+1} \,|\, A_N, B_N, C_N, D_N, \mathcal{D}) =$$

$$\sum_{k=1}^{256} P(B_{N+1}, C_{N+1} \,|\, A_N, B_N, C_N, D_N, \text{Rule}_k) \, P(\text{Rule}_k \,|\, \mathcal{D}) \quad (16.1)$$

As things stand in full generality, the IP would be forced to construct a 64 cell joint probability table for this problem. However, if we use the **Product Rule** on the first term on the *rhs*, it breaks down into a product of two probabilities,

$$P(B_{N+1}, C_{N+1} \,|\, A_N, B_N, C_N, D_N, \text{Rule}_k) =$$

$$P(B_{N+1} \,|\, C_{N+1}, A_N, B_N, C_N, D_N, \text{Rule}_k) \times P(C_{N+1} \,|\, A_N, B_N, C_N, D_N, \text{Rule}_k) \quad (16.2)$$

The rules of the game specify that B_{N+1} does not depend on C_{N+1} or D_N, while C_{N+1} does not depend on A_N. Thus, we can simplify the first term on the *rhs* of Equation (16.1) to,

$$P(B_{N+1}, C_{N+1} \,|\, A_N, B_N, C_N, D_N, \text{Rule}_k) =$$

$$P(B_{N+1} \,|\, A_N, B_N, C_N, \text{Rule}_k) \times P(C_{N+1} \,|\, B_N, C_N, D_N, \text{Rule}_k) \quad (16.3)$$

Thus, by relying on this kind of conditional independence, we can escape the tiresome labor of constructing a 64 cell table. We can get by with our original 16 cell joint probability table with only one extra multiplication.

Recall that, fortunately, we have already segregated the impact of the data to the second term. None of the data have any impact on the first term; the data's role is to modify the relative standing of all the models we started out with. Instead of averaging over all the models, with every such model possessing an equal weight, which we had been forced to do under "total ignorance," the data will permit us to reorder the relative status of the models.

These last few examples are a parting nod to the formal manipulation rules on which we placed such emphasis in this beginning Volume. There is really a lot to be gained from these rules in exchange for a minimal expenditure of mathematical complexity.

The emphasis will shift drastically away from these formal manipulation rules in Volume II. There, the focus is on finding a good algorithm for making the numerical assignments that are somehow in consonance with the information inherent in some model.

To exhibit a workable numerical example of conditional independence, limit the space of rules to just nine of the deterministic models that are available from the full set of 256 rules comprising the ECA. Suppose again that one data point was observed from an unknown ontological system running according to one of these deterministic rules. The updated color of cell B at the second time step was black when at the beginning time step it was white and its two neighbors cells were black. Write this data point as,

$$\mathcal{D} \equiv \{B_2 = b, A_1 = b, B_1 = w, C_1 = b\}$$

What is the probability that the system will color both cells B and C black at time step $N + 1$ given that cells A, B, C, and D at time step N were, respectively, white, black, white, and black? This state of knowledge is also conditioned on the given data.

As shown in the series of Exercises 16.9.5 through 16.9.13, the answer is,

$$P(B_{N+1} = b, C_{N+1} = b \mid A_N = w, B_N = b, C_N = w, D_N = b, \mathcal{D}) = 4/5$$

The IP's state of knowledge that the statement, "Cells B_{N+1} and C_{N+1} will both be colored black." is, in fact, TRUE is 80% on the quantitative scale represented by probability measure. There are three other possibilities for the colors of the updated cells. Their combined probability must be only 1/5.

16.5   Growth of the Joint Probability Table

If we want CA to stand in for arbitrarily complicated ontological systems, then these elementary cellular automata we have been looking at can be generalized to depend on many different colors of many neighboring cells existing arbitrarily back into the past. We hinted at this generality at the end of Chapter Three when we examined a Boolean function with five arguments, and a carrier set consisting of four elements.

Thus, a more complicated CA might look at *two* neighbors to the right and left of the previous cell, instead of a single neighbor, to update its color. That color might now be light gray, dark gray, black, or white instead of just black or white.

In Chapter Fourteen, college success depended upon three causal factors, namely, the three test scores. Similarly, the updated color of the cell in the ECA depended upon the colors of three cells at the past time step. The example with the college students was different to the extent that the causal factors were taken at the same time (on the same student) rather than existing one time step in the past.

Prediction problems can be generalized to many statements that need to be predicted while depending upon many causal factors and some data. For example, predicting a student's graduation *and* job success could be seen to depend on 5 tests

with each test's score broken down into quartiles, instead of medians. Presumably, past data consisting of an accurate recording of graduation, job success, and the test scores for N students also exist.

We could write down the prediction equation for a more complicated CA system with two neighbors while still retaining just two colors as,

$$P(C_{N+1}, D_{N+1} \,|\, A_N, B_N, C_N, D_N, E_N, F_N, \mathcal{D}) =$$

$$\sum_{k=1}^{\mathcal{M}} P(C_{N+1}, D_{N+1} \,|\, A_N, B_N, C_N, D_N, E_N, F_N, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, \mathcal{D})$$

We can generalize easily enough in a formal sense by writing out equations like this, but we quickly become aware of some practical computational problems. We can still employ conditional independence so that the first term can be broken down into a multiplication of two terms depending on a smaller joint probability table. But even with conditional independence, we are forced to construct a 64 cell joint probability table so that various models could fill in their numerical assignments to the joint statements.

What if we wanted to look at a wide range of student characteristics represented symbolically by $A, B, C, \cdots$, and thought to be influenced by a potentially large spectrum of causal factors, $\cdots X, Y, Z$? Although the formal generalization of the prediction equation for the next student's characteristics, $A_{N+1}, B_{N+1}, \cdots$, after having observed the data from N students, can be written out as,

$$P(A_{N+1}, B_{N+1}, C_{N+1}, \cdots \,|\, \cdots, X_{N+1}, Y_{N+1}, Z_{N+1}, \mathcal{D}) =$$

$$\sum_{k=1}^{\mathcal{M}} P(A_{N+1}, B_{N+1}, C_{N+1}, \cdots \,|\, \cdots, X_{N+1}, Y_{N+1}, Z_{N+1}, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, \mathcal{D})$$

we see that the joint probability table required to solve Bayes's Theorem in the first term on the *rhs* would be absolutely enormous,

$$P(A, B, C, \cdots \,|\, \cdots, X, Y, Z, \mathcal{M}_k) = \frac{P(A, B, C, \cdots, X, Y, Z \,|\, \mathcal{M}_k)}{P(\cdots X, Y, Z \,|\, \mathcal{M}_k)}$$

As a thought experiment, suppose that we want to predict the color of all the cells in the CA under some model arbitrarily far into the future. Thus, we want not only to assess a state of knowledge about B_{N+1}, but also about C_{N+2}, D_{N+3}, all the way up to Y_{N+M}. Here, Y_{N+M} is simply a way of denoting the color of, say, the 10,000th cell at some time step M, say, at the 100,000th time step, far into the future.

$$P(B_{N+1}, C_{N+2}, \cdots, Y_{N+M} \,|\, A_N, B_N, C_N, \cdots Z_N, \mathcal{M}_k) =$$

$$\frac{P(B_{N+1}, C_{N+2}, \cdots, Y_{N+M}, A_N, B_N, C_N, \cdots Z_N, \,|\, \mathcal{M}_k)}{P(A_N, B_N, C_N, \cdots, Z_N, \,|\, \mathcal{M}_k)}$$

Think of the absolutely unfathomable gigantic joint probability table that ensues for the numerator in Bayes's Theorem, together with the required summation over selected cells in this table to find the denominator. Indeed, merely taking notice of the requirement to allocate some probability to all of the statements in this inconceivably large state space indicates the futility of the enterprise.

From this probabilistic perspective, we gain an even greater appreciation for Wolfram's insistence on not being able to predict the detailed behavior of the cellular automaton far into the future.

16.6   Inferences About Other Statements

Up till now we have been talking mainly about how inferencing and probability handle an IP's ignorance (or knowledge if you prefer) of the rules that run the ECA. But inferencing is completely general. We can move statements around, or include or not include them within the probability operator, as we please. We'll set up a problem in the context of an ECA, but use the simpler notation of logic functions.

We have been making inferences in order to predict the color of a cell of an ECA that needs to be updated at the next time step. The IP was assumed "ignorant" about the actual rule running the ECA, so the prediction had to be averaged over the many predictions from however many rules were under consideration. If there were some data, then these data could revise the relative standing of the rules from their initial uniform status. Now, for a change of pace, suppose that the rule is, in fact, known, but a neighbor, and say it is A_N, is not known.

To warm up, ponder this argument from a non–probabilistic viewpoint. We find ourselves in the domain of Classical Logic thinking about the implication function, $A \to B$. Or, $f_{13}(A, B) = AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$. If you are told that $B = T$ and the output of the logic function is also T, then you isolate the two cases fitting this description, $f_{13}(A = T, B = T) = T$ and $f_{13}(A = F, B = T) = T$. In one case, $A = T$ and the other $A = F$, so you argue that your state of knowledge about A is accurately reflected by the number $1/2$.

In probabilistic notation we think about this as,

$$
\begin{aligned}
P(A = T \mid B = T, Z = T, A \to B) &= \frac{P(ZAB \mid \mathcal{M}_k)}{P(ZB \mid \mathcal{M}_k)} \\[2ex]
&= \frac{P(ZAB \mid \mathcal{M}_k)}{P(ZAB \mid \mathcal{M}_k) + P(Z\overline{A}B \mid \mathcal{M}_k)} \\[2ex]
&= \frac{1/4}{1/4 + 1/4} \\[2ex]
&= 1/2
\end{aligned}
$$

The IP wanted to make an inference about A. Therefore, it was placed to the left of the conditioned upon symbol.[1] Both B and the functional assignment Z were known, so they were placed to the right of the solidus. The rule, or particular logic function, or model $\mathcal{M}_k$ was also assumed to be given, and appeared to the right of the solidus as well.

The formal manipulation rule known as Bayes's Theorem takes this set up of the knowns and unknowns and transforms it into probabilities for joint statements of everything being considered. Contingent on some legitimate numerical assignment to the abstract probabilities appearing in the numerator and denominator, (and this would come about through the information from model $\mathcal{M}_k$), a correct state of knowledge is returned to the IP.

Do the same thing for a three variable Boolean function, $f_{110}(A, B, C)$. A is not known, but B and C are known as well as the fact that Rule 110 is controlling the ontological system. What is the IP's state of knowledge that the left neighbor A_N is white given that the cell was black at the previous time step, and its right neighbor C_N was white? The updated cell color was black. The IP assesses its state of knowledge about A as,

$$P(A = F \mid B = T, C = F, Z = T, \text{Rule 110}) = \frac{P(Z\overline{A}B\overline{C} \mid \text{Rule 110})}{P(ZB\overline{C} \mid \text{Rule 110})}$$

$$= \frac{P(Z\overline{A}B\overline{C} \mid \text{Rule 110})}{P(Z\overline{A}B\overline{C} \mid \text{Rule 110}) + P(ZAB\overline{C} \mid \text{Rule 110})}$$

$$= \frac{\text{Cell 6}}{\text{Cell 6} + \text{Cell 2}}$$

$$= \frac{1/8}{1/8 + 1/8}$$

$$= 1/2$$

The joint probability table for Rule 110 that first appeared in Chapter Nine (Please refer back to Figure 9.1 on page 224) was consulted in order to fill in the "correct" numerical assignments to the probabilities for the joint statements appearing in the numerator and denominator. These numerical assignments were "correct" in the sense of incorporating the required information under the specified model for Rule 110.

Examine Rule 110 for the two cases when the updated cell is black when it was black at the previous time step and the right neighbor was white. In one case, the left neighbor is white, and in the other case the left neighbor is black.

---

[1] Also known as the *solidus*.

16.7   A Discomforting Realization

Now for the *coup de grâce*. When everything is specified or known, we have a *deduction*. All computations take place without the need for any probabilities to appear. We have a deterministic ontological system.

If the starting configuration of a CA is specified, if the rule running the CA is specified, and if the number of time steps is specified, the *Mathematica* function `CellularAutomaton[` ⋯ `]` proceeds to calculate the millionth time step of some CA without ever resorting to anything called probability or inferencing.

On the other hand, if some things are not known or specified, an IP may not be able to *deduce* anything. Computer programs do not gracefully degrade when you do not provide the proper input. The IP has no recourse but to fall back on probability and inferencing.

Recall the inability of an IP to logically deduce anything about A in the case of an implication, $A \rightarrow B$, even when B was known. Contrary to this state of paralysis, an IP at least ends up in some state of knowledge after an inference. Such a state of knowledge will indicate to the IP just how strongly it may believe that some statement is TRUE.

The initial pessimistic conclusion from the few numerical examples we have conducted so far is that it is terribly difficult to predict the future in all of its fine detail. The combinatorial explosion that mandates carving up the probability over myriad possibilities renders moot any state of knowledge about the details of an ontological system into the far future.

Unfortunately, neither can the IP predict the far future in the ideal case of a deductive deterministic system emulating a complicated ontology. All it can do is compute as fast as its little feet can run. Wolfram was pessimistic that any IP could ever compute faster than the Universe itself was computing reality. Thus, if an IP can never win this race, its hope for predicting the evolution of an ontological system seems doomed.

16.8   How to Proceed?

If there were anything to be learned from these exercises, it was the realization of how fiendishly difficult it was to solve the prediction equation for even the simplest of these abstract ontological systems. This lends weight to Wolfram's repeated injunction that the only way to find out what a CA is going to do in the far future is to simply let it run for some required number of steps.

Inferencing problems that fall under the same template of our college graduation scenario are vastly easier than predicting the CA. Predicting success for each new student, no matter how many new students were envisioned, depended only on,

say, the three or more test scores for that student. Thus, the joint probability tables, the required computations using Bayes's Theorem, and even averaging over many models all seemed feasible. The universe of all possible influences on college graduation was whittled down so that we could handle the problem computationally.

Statistical science has made great strides in the last few decades in trying to incorporate increasingly complex models of many interacting variables. Various sophisticated versions of conditional independence have been tried in order to reduce the size of unadulterated joint probability tables.

Various techniques that travel under the labels of *Hidden Markov Models* and *Linear Dynamic Models* have tried to come to grips with predicting observations, connected in time, into the far future. Again, these techniques rely on clever ways of reducing the dependence of what the IP is trying to predict from everything that has happened prior to that point.

Ideally, if a small set of causal factors determines an observation at each time step, then the models don't have to take into account all of the past observations. Furthermore, it then becomes feasible to imagine that this small set of causal factors might itself be slowly temporally evolving.

The evolution of the CA depends pretty rigidly upon the history of everything that has happened in its world back to the very beginning. The IP can't carve off a manageable chunk of the world, and find restricted computations that depend upon this limited set of causal factors. The CA are models of a world where everything that has ever happened in the world is interconnected both spatially and temporally. Combinatorial explosion sets in early and with a vengeance.

However, we are not going to give up quite yet. We are ready to concede Wolfram's point that the computational effort in predicting the fine details of the CA arbitrarily far into the future is more than daunting. Nonetheless, if we remember the lessons learned in generalizing Classical Logic, we might hold out hope for some kind of generalized inferential process that might predict something interesting during the evolution of the CA.

Some way has to be found of short–circuiting the overwhelming computational effort outlined above in predicting the far future. There *is* more computation in an inferential system where certain things are not known than in a deterministic system where everything is specified. By making things seemingly more complicated, could it be that we actually find our way to a solution?

Within deductive schemata such as Classical Logic, we may find undecidable statements staring us in the face. But we tried to show that when we generalized the deductive process with probability and made it inferential, formerly undecidable statements were transformed into probabilistic statements. If we generalize the elementary CA, essentially what we do is to allow for a full range of legitimate numerical assignments into the joint probability table. This implies that we no longer have a deductive deterministic process, but rather an inferential process.

But ultimately something has to give. There has to be some trade–off that allows an inferential approach to succeed where a deductive approach fails. Is it in abandoning certain fine–level details that don't end up making a difference anyway in some final conclusion?

The analogy here is to the same reality that had to be faced historically by statistical mechanics and thermodynamics. It quickly became clear that it would be impossible to follow the trajectory of every atom in a gas even though the underlying physics governing the dynamics of each atom, and its interaction with other atoms, was known.

An inferential process that dealt with large scale macroscopic averages of the fundamental atomic dynamics was adopted; the deterministic fine scale microscopic information about atomic dynamics simply had to be abandoned.

In a more speculative mode, one would concede that any given human experience must be constructed from some untold number of different details of the quantum states of our neural and hormonal systems. Suppose I am listening to Beethoven's Piano Concerto Number 5, and this activity is accompanied by a certain pleasurable feeling. Every time I listen to this piece of music, and experience roughly the same agreeable auditory effect, the quantum details in my brain underlying this experience must be vastly different. But apparently none of those details make any difference to me as far as my own subjective appreciation of this piece of music.

Likewise, if we were to abandon the quest of stubbornly trying to follow the detailed microscopic features of an evolving CA, and concentrate rather on some large grained macroscopic structures that are probabilistic averages, then we might hope to find some computational solution to circumvent the prediction problem.

16.9   Solved Exercises for Chapter Sixteen

Exercise 16.9.1. Find the probability for the functional assignment given the two arguments of a logic function. The IP is completely uninformed about which logic function applies.

Solution to Exercise 16.9.1

This is the solution to the problem posed in section 16.2.

$$P(Z = T \,|\, A = T, B = F) = \sum_{j=1}^{16} P(Z = T \,|\, A = T, B = F, f_j \,[\, A, B \,]\,) \; P(f_j \,[\, A, B \,]\,)$$

The first term on the right hand side will be found using Bayes's Theorem. The numerator in Bayes's Theorem will look like $P(ZA\overline{B} \,|\, f_j \,[\, A, B \,]\,)$. $ZA\overline{B}$ refers to the joint statement in cell 3 of the eight cell joint probability table. The denominator in Bayes's Theorem will be the sum of the probability of this term, and the probability for the joint statement $\overline{Z}A\overline{B}$ in cell 7 of the joint probability table. Thus, Bayes's Theorem calculates the first term as,

$$P(Z = T \,|\, A = T, B = F, f_j \,[\, A, B \,]\,) = \frac{\text{Cell 3}}{\text{Cell 3} + \text{Cell 7}}$$

Determine which of the 16 logic functions contain the term $A\overline{B}$ in their DNF expansion. Those functions possessing this term will have some positive probability assigned to cell 3 and zero assigned to cell 7. Thus, the probability that $Z = T$ will be 1 for these functions. Eight logic functions, $f_4, f_7, f_9, f_{11}, f_{12}, f_{14}, f_{15}$, and f_{16} have the $A\overline{B}$ term.

The functions which do not have the $A\overline{B}$ term in their DNF expansions will have a zero assigned to cell 3. Therefore, all of these probabilities will be zero. There are eight functions, $f_1, f_2, f_3, f_5, f_6, f_8, f_{10}$, and f_{13} which fall into this category.

After assigning an equal probability of $P(f_j \,[\, A, B \,]\,) \equiv P(\mathcal{M}_k) = \frac{1}{16}$ as the prior probability for each function, we find that,

$$P(Z \,|\, A, \overline{B}) = \frac{1}{16} \,(0 + 0 + 0 + 1 + 0 + 0 + 1 + 0 + 1 + 0 + 1 + 1 + 0 + 1 + 1 + 1) = 1/2$$

This result makes eminent intuitive sense.

Exercise 16.9.2. Find the probability for the functional assignment given the two arguments of a logic function just as in the first exercise. The IP however has observed the functional output once.

Solution to Exercise 16.9.2

This is the solution to the problem in section 16.2 when the IP was not completely uninformed, but now has in its possession one data point. It is best to apply the rules in a strictly mechanical fashion to find the result. Then, one can go back and find other alternative explanations that are perhaps more intuitively satisfying.

The functional output was observed to be T when $A = T$ and $B = T$. In other words, the one data point is,

$$\mathcal{D} = \{Z_1 = T, A_1 = T, B_1 = T\}$$

The second term on the right hand side of the prediction equation is expanded by Bayes's Theorem into,

$$P(f_j\,[\,A, B\,]\,|\,\mathcal{D}) = \frac{P(\mathcal{D}\,|\,f_j\,[\,A, B\,])\,P(f_j\,[\,A, B\,])}{\sum_{j=1}^{16} P(\mathcal{D}\,|\,f_j\,[\,A, B\,])\,P(f_j\,[\,A, B\,])}$$

Just as before, there are going to be eight logic functions which have an AB term in the DNF expansion. Therefore, these eight functions will be certain to output a T when the arguments are $A = T$ and $B = T$. These eight functions are $f_5, f_8, f_{10}, f_{11}, f_{13}, f_{14}, f_{15}$, and f_{16}.

The first term in the numerator will then be 1 for these eight functions, and 0 for the remaining eight functions. The initial probability for the functions is still $1/16$. The sum in the denominator will then be $8/16$, and the numerator will either be 0 or $1/16$ for any given function. Thus, we have a probability of 0 or $1/8$ for the probability of a function conditioned on the one piece of data, $P(f_j\,[\,A, B\,]\,|\,\mathcal{D})$.

The first term in the overall prediction equation,

$$P\,[\,(Z = T\,|\,A = T, B = F, f_j(A, B)\,]$$

is exactly the same as it was before. Only the probabilities for the functions, the second term in the prediction equation, have been modified because of the presence of the data. They are not all equal to $1/16$ when the IP was totally ignorant.

Eight models have a probability of 0 because of the data. They have been ruled out of contention. The eight models that remain in contention each have a probability of $1/8$. We have reached the final answer that,

$$P(Z_2 = T\,|\,A_2 = T, B_2 = F, \mathcal{D}) =$$

$$\sum_{j=1}^{16} P(Z_2 = T\,|\,A_2 = T, B_2 = F, f_j\,[\,A, B\,])\,P(f_j\,[\,A, B\,]\,|\,\mathcal{D}) = 1/2$$

We found out in the first exercise that only functions $f_4, f_7, f_9, f_{11}, f_{12}, f_{14}, f_{15}$, and f_{16} were certain to output a T given that $A = T$ and $B = F$. But now, after the data, four of these functions, f_4, f_7, f_9, and f_{12}, have been ruled out. They have a probability of 0. The probability of the function to output T, which is 1, from the four logic functions f_{11}, f_{14}, f_{15}, and f_{16} that have not been ruled out by the data are weighted in the averaging procedure by 1/8.

However, the other four logic functions that remained in contention after the datum was observed, f_5, f_8, f_{10}, and f_{13} have a probability of 0 of outputting a T given that $A = T$ and $B = F$. Thus,

$$P(Z_2 \,|\, A_2, \overline{B}_2, \mathcal{D}) = 1/8 \,[\, 0 + 0 + 0 + 1 + 0 + 1 + 1 + 1 \,] = 1/2$$

Exercise 16.9.3. Assess the state of knowledge about a second student's graduation success given the data from the first student in the context of the last exercise.

Solution to Exercise 16.9.3

Within the context of section 16.2 and Expression 3, a student takes two tests. Suppose we restrict ourselves to just sixteen models that assign numerical values to the eight cells of the joint probability table. This is done in order to match up with the 16 logic functions.

What is the probability that the next student graduates given that he or she scored HIGH on the first test and LOW on the second? The IP knows the result that the first student graduated after scoring HIGH on both tests.

In terms of the notation for graduation success, write out the prediction equation for the second student as,

$$P(A_2 = G \,|\, B_2 = H, C_2 = L, \mathcal{D}) =$$

$$\sum_{k=1}^{16} P(A_2 = G \,|\, B_2 = H, C_2 = L, \mathcal{M}_k) \, P(\mathcal{M}_k \,|\, \mathcal{D})$$

Let's examine two of the models in detail. From the overall total of sixteen models, select model $\mathcal{M}_{11}$ and model $\mathcal{M}_{10}$. These two models correspond to logic function $f_{11}(A, B)$, $A \lhd B$, and logic function $f_{10}(A, B)$, $A \rhd B$. $f_{11}(A, B)$ has the DNF expansion of $AB \lor A\overline{B}$, while $f_{10}(A, B)$ has the DNF expansion of $AB \lor \overline{A}B$.

From the DNF expansions, we observe that both models have a coefficient of T for AB. This is why both of these models, together with six other models, were supported by the one data point where $\mathcal{D} \equiv \{A = G, B = H, C = H\}$.

However, the IP seeks to assess its state of knowledge about $A = G$ when $B = H$ and $C = L$. The DNF expansion of model $\mathcal{M}_{11}$ also has a coefficient of T for $A\overline{B}$, or $B = H$ and $C = L$. Thus, this particular model outputs a probability of 1 with a weight of 1/8 from the reorientation of the model space dictated by the one data point.

On the other hand, the DNF expansion of model $\mathcal{M}_{10}$ has a coefficient of F for $A\overline{B}$. Thus, this particular model outputs a probability of 0 with a weight of $1/8$ from the reorientation of the model space dictated by the one data point.

The other six models selected by the data point have the same kind of symmetry. This leads to the result that the probability of graduation for the second student is still only $1/2$ given these particular test scores, and despite the increased knowledge about the model space.

Exercise 16.9.4. Validate the argument made in the last exercise with a calculation based on the joint probability tables for these two models and Bayes's Theorem.

Solution to Exercise 16.9.4

The joint probability table will have as many cells as the dimension of the state space, and for this problem this means that $n = 8$.

The DNF expansion for model $\mathcal{M}_{11}$ tells us that this model will insert the information that cells 1 and 3 will have a numerical assignment of $1/4$ with cells 2 and 4 having assignments of 0. Likewise, cells 6 and 8 will have a numerical assignment of $1/4$, and cells 5 and 7 an assignment of 0.

Bayes's Theorem tells us that conditioned on model $\mathcal{M}_{11}$, the probability of graduation with a HIGH score on test 1 and a LOW score on test 2 is,

$$P(G \,|\, H, L, \mathcal{M}_{11}) \;=\; \frac{P(G, H, L \,|\, \mathcal{M}_{11})}{P(H, L \,|\, \mathcal{M}_{11})}$$

$$=\; \frac{\text{Cell } 3}{\text{Cell } 3 + \text{Cell } 7}$$

$$=\; \frac{1/4}{1/4 + 0}$$

$$=\; 1$$

Everything works in the same way for model $\mathcal{M}_{10}$. The DNF expansion for model $\mathcal{M}_{10}$ tells us that this model will insert the information that cells 1 and 2 will have a numerical assignment of $1/4$ with cells 3 and 4 having assignments of 0. Likewise, cells 7 and 8 will have a numerical assignment of $1/4$, and cells 5 and 6 an assignment of 0.

Bayes's Theorem tells us that conditioned on model $\mathcal{M}_{10}$, the probability of graduation with a HIGH score on test 1 and a LOW score on test 2 is,

$$P(G \,|\, H, L, \mathcal{M}_{10}) \;=\; \frac{P(G, H, L \,|\, \mathcal{M}_{10})}{P(G, H, L \,|\, \mathcal{M}_{10}) + P(\overline{G}, H, L \,|\, \mathcal{M}_{10})}$$

$$=\; \frac{0}{0 + 1/4}$$

$$=\; 0$$

One model predicts that it is certain the student will graduate, while the other model insists with just as much certainty that the student will not graduate. Since both models have equal weight after the observation on the first student, the state of knowledge conditioned on the two causal factors is no different than the marginal probability for graduation (after going through the same exercise for the other six models).

Both models accurately predicted that a student with HIGH scores on both tests would graduate, so their status *vis-à-vis* the one data point remained the same as when they started. A model like model $\mathcal{M}_9$, based on logic function $A \oplus B$, predicted with certainty that a student with HIGH scores on both tests would not graduate. Our one data point invalidated such a model.

Thus, this model's prediction for the second student, whatever it might be, (certainty that the student will graduate) is not counted. It is like the model that predicts a trick coin with two HEADS that is immediately invalidated when the first flip shows TAILS.

Exercise 16.9.5. Examine the DNF expansions of nine ECA rules. Start by picking a rule with no basis functions, and end with a rule with all eight basis functions in the DNF expansion.

Solution to Exercise 16.9.5

With this exercise, we begin the solution of the problem posed in section 16.4. The completion of the problem takes place at Exercise 16.9.13.

We'll dispose of three rules quickly. The DNF expansion of a Boolean function with no terms, that is, the coefficient of every basis function is F, is the function that outputs F for all arguments. This is Rule 0 which outputs all white cells for any starting configuration.

The DNF expansion of a Boolean function consisting of all eight terms, that is, where the coefficient of every basis function is T, is the function that outputs T for all arguments. This is Rule 255 which outputs all black cells for any starting configuration.

We'll pick Rule 110 for the Boolean function with five terms in the expansion. Now, with these three rules selected, we may arbitrarily choose six other rules with one, two, three, four, six, and seven terms in their DNF expansions.

Next, pick the rule with the basis function ABC with coefficient T in the DNF expansion for the next rule with one term. This is Rule 128. Refer back to Wolfram's numbering scheme in section 3.3.1 to refresh your memory as to how to find this rule number.

This Boolean function of three variables has a functional assignment of T for arguments $A = T, B = T$, and $C = T$. Thus, there is a T in the second row under TTT, and an F for the other seven columns. Table 16.1 is just like the other functional assignment tables shown in Chapter Three.

Table 16.1: *The functional assignment which is Rule 128.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
T	F	F	F	F	F	F	F

The binary number for this Boolean function is,

$$(1 \times 2^7) + (0 \times 2^6) + (0 \times 2^5) + (0 \times 2^4) + (0 \times 2^3) + (0 \times 2^2) + (0 \times 2^1) + (0 \times 2^0) = 128$$

This rule outputs a black cell if the three previous cells were all black. In all other cases, the cell is colored white.

For a rule with two terms, pick $ABC \vee AB\overline{C}$. This is Rule 192. There is a T under the first two columns, so the rule number is,

$$(1 \times 2^7) + (1 \times 2^6) + (0 \times 2^5) + (0 \times 2^4) + (0 \times 2^3) + (0 \times 2^2) + (0 \times 2^1) + (0 \times 2^0) = 192$$

This rule outputs a black cell if the three previous cells were all black, or if the cells were black, black, and white. In all other cases, this ontological system produces a white cell. This is also a Boolean function of three variables which we examined before in Chapter Three.

Continue in the same manner to pick the remaining four rules. This is all summarized below in Table 16.2.

Table 16.2: *Summary of nine ECA rules used as models in the numerical example.*

k	Model k	Full DNF Expansion	Binary
1	0	F	00000000
2	128	ABC	10000000
3	192	$ABC \vee AB\overline{C}$	11000000
4	56	$A\overline{B}C \vee A\overline{B}\,\overline{C} \vee \overline{A}BC$	00111000
5	85	$AB\overline{C} \vee A\overline{B}\,\overline{C} \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}$	01010101
6	110	$AB\overline{C} \vee A\overline{B}C \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$	01101110
7	126	$AB\overline{C} \vee A\overline{B}C \vee A\overline{B}\,\overline{C} \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$	01111110
8	254	$ABC \vee AB\overline{C} \vee A\overline{B}C \vee A\overline{B}\,\overline{C} \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$	11111110
9	255	T	11111111

Exercise 16.9.6. Double–check that Model 5's binary number is correct.

Solution to Exercise 16.9.6

What we have called Model 5 for the numerical example is really Rule 85 for the ECA. Deconstructing the binary number 01010101 reveals that,

$$64 + 16 + 4 + 1 = 85$$

The pattern of 1s and 0s guides us in forming Table 16.3 to assist in the DNF expansion. Thus, including the orthonormal basis functions with a coefficient of T,

Table 16.3: *The functional assignment which is Rule 85.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
F	T	F	T	F	T	F	T

and excluding the basis functions with a coefficient of F, we have,

$$f_{85}(A, B, C) = AB\overline{C} \vee A\overline{B}\,\overline{C} \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}$$

Exercise 16.9.7. Can the full DNF expansions be simplified?

Solution to Exercise 16.9.7

Yes, they can. We have investigated how the full DNF expansion for Rule 110 can be simplified. Rule 255 possess all eight orthonormal basis functions, and it simplifies to T. Look at Rule 85 again whose full DNF expansion was found in the previous exercise,

$$f_{85}(A, B, C) = AB\overline{C} \vee A\overline{B}\,\overline{C} \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}$$

This expression can be simplified using the formal manipulation rules applicable to Boolean functions. Rule 85 actually reduces to,

$$f_{85}(A, B, C) = \overline{C}$$

In other words, the color of the updated cell B_{N+1} is determined by just looking at the right neighbor C_N, and taking the opposite color.

Here is a theorem, built up step by primitive step from the axioms, proving that Rule 85 is $\overline{C}$. Factor out the common term $\overline{C}$ from the full DNF expression,

$$AB\overline{C} \vee A\overline{B}\,\overline{C} \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C} = \overline{C} \wedge (AB \vee A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B})$$

We know we are on the right track at this point because we have seen the expression in parentheses before as the decomposition of a joint probability table where the

four terms are the four cells that must add up to 1. Continuing with the formal manipulations,

$$(AB \vee A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B}) = (A \wedge (B \vee \overline{B})) \vee (\overline{A} \wedge (B \vee \overline{B}))$$

$$(A \wedge (B \vee \overline{B})) \vee (\overline{A} \wedge (B \vee \overline{B})) = (A \wedge T) \vee (\overline{A} \wedge T)$$

$$(A \wedge T) \vee (\overline{A} \wedge T) = A \vee \overline{A}$$

$$A \vee \overline{A} = T$$

$$\overline{C} \wedge T = \overline{C} \qquad\qquad \text{QED}$$

Exercise 16.9.8. Given the data, reorder the relative standing of all nine models.

Solution to Exercise 16.9.8

The datum that was obtained from the ontological system indicated that cell B_{N+1} was black when A_N and C_N were black and B_N was white. Look at all nine models to see which ones violate this datum. These models will be excluded in the overall averaging process.

Rules 0, 128, 192, and 85 do not output a black cell when the causal factors are black, white, and black. These four rules are therefore immediately rejected on the basis of a measurement on the system, leaving five rules still in the running. These five rules, Rules 56, 110, 126, 254, and 255, do output a black cell given the described status of the causal factors. Their predictions must be included in the overall averaging process over the model space.

Formally, we have the revised probability for the models as conditioned on the data,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}$$

The IP began in a state of total ignorance, so $P(\mathcal{M}_k) = 1/9$. The probability of the black cell from these deterministic models, $P(\mathcal{D} \,|\, \mathcal{M}_k)$, will be either 1 or 0. The numerator will be either a 0 or 1/9. As already mentioned, only five of the rules output a black cell, so the denominator will contain the sum 5/9. The revised probability for the four excluded rules will now be 0, and the revised probability for the five retained rules will be 1/5.

Exercise 16.9.9. Draw a sketch for Rule 85 showing that it cannot output a black cell under the stated conditions.

Solution to Exercise 16.9.9

From Table 16.3 in Exercise 16.9.6, we note that the Boolean function implementing Rule 85 has a functional assignment of F when its arguments are T, F, and T,

$$f_{85}(A = T, B = F, C = T) = F$$

Thus, this rule must output a white cell when its causal factors are the stated colors, black, white, and black, of the neighbor cells. It cannot output a black cell.

Here is a sketch in Figure 16.1 of a portion of the elementary cellular automaton running under Rule 85 (Model 5 in our example) as it evolves into the next time step.

Figure 16.1: *An ontological system (represented by a cellular automaton) running according to one of the deterministic rules in the numerical example, Rule 85.*

Exercise 16.9.10. Does this discussion about the data have any direct impact on the first term in the prediction equation?

Solution to Exercise 16.9.10

No, it does not. What the IP is interested in predicting is whether the two cells B_{N+1} and C_{N+1} will be black based on differing causal factors in the ontological system than appeared in the data. So it must still compute the probabilities under all models in the first term.

Now, just as before, these probabilities must be either 1 or 0 because they arise from a deterministic system. A probability of 1, however, might come from a model that was ruled out by the data. Or a probability of 0 might arise from a model that was retained. In either case, the net result is a 0 in the overall summation. Only those models that were retained, and which have a probability of 1 for the new settings of the causal factors, will add a 1 to the overall summation.

Of course, the data play a vital role in reordering the original standings of the models. So even though a model might produce the desired output for the new settings of the causal factors, if it were ruled out by that first data point, it has no further influence on the ultimate assessment of the IP's state of knowledge.

Exercise 16.9.11. Show the details of the prediction equation in the solution of the numerical example.

Solution to Exercise 16.9.11

Begin this task by breaking down the overall equation into manageable chunks through conditional independence. We seek the probability that the two cells to be updated will both be black. We know the color of the relevant cells at the previous time step. We know that sometime in the past, the system output black when the causal factors were black, white, and black. Encapsulate all of this within the left hand side of the prediction equation,

$$\text{lhs} \equiv P(B_{N+1} = b, C_{N+1} = b \,|\, A_N = w, B_N = b, C_N = w, D_N = b, \mathcal{D} = \{b, b, w, b\})$$

Simplifying and invoking Equations (16.2) and (16.3),

$$\text{lhs} = \sum_{k=1}^{9} P(B_{N+1}, C_{N+1} \,|\, \overline{A}_N, B_N, \overline{C}_N, D_N, \mathcal{M}_k) \times P(\mathcal{M}_k \,|\, \mathcal{D})$$

$$= \sum_{k=1}^{9} P(B_{N+1} \,|\, \overline{A}_N, B_N, \overline{C}_N, \mathcal{M}_k) \times P(C_{N+1} \,|\, B_N, \overline{C}_N, D_N, \mathcal{M}_k) \times P(\mathcal{M}_k \,|\, \mathcal{D})$$

There will be nine terms in the summation. Each one of these terms will consist of three entries that are multiplied. The first entry is the probability under the kth model for the first updated cell to be black. The second entry is the probability under the kth model for the second updated cell to be black. The third entry is the revised probability of that k^{th} model after the data have been processed.

So the first term in the summation will look like $(0 \times 0 \times 0)$ because Rule 0 ($k = 1$) has zero probability of outputting a black cell under the stated conditions for both updated cells. In addition, Rule 0 was eliminated on the basis of the one piece of data because it could not output a white cell when the causal factors were black, white, and black.

The fourth term will look like $(0 \times 1 \times 1/5)$ because Rule 56 ($k = 4$) has zero probability of outputting a black cell for B_{N+1}. But it does have probability 1 for outputting a black cell for C_{N+1}. Also, Rule 56 was one of the rules that was retained because the data did not eliminate it. It has a revised probability of $1/5$.

The sixth term will look like $(1 \times 1 \times 1/5)$ because Rule 110 ($k = 6$) has a probability of 1 for outputting a black cell for both updated cells under the stated conditions. It also was kept as a rule, and its revised probability is $1/5$ after the data were observed.

Explicitly, the sum looks like,

$$\underbrace{(0 \times 0 \times 0)}_{\text{Rule 0}} + \underbrace{(0 \times 0 \times 0)}_{\text{Rule 128}} + \underbrace{(0 \times 0 \times 0)}_{\text{Rule 192}} + \underbrace{(0 \times 1 \times 1/5)}_{\text{Rule 56}} + \underbrace{(1 \times 0 \times 0)}_{\text{Rule 85}} +$$

$$\underbrace{(1 \times 1 \times 1/5)}_{\text{Rule 110}} + \underbrace{(1 \times 1 \times 1/5)}_{\text{Rule 126}} + \underbrace{(1 \times 1 \times 1/5)}_{\text{Rule 254}} + \underbrace{(1 \times 1 \times 1/5)}_{\text{Rule 255}}$$

In this way, we have arrived at the final answer of,

$$P(B_{N+1} = b, C_{N+1} = b \mid A_N = w, B_N = b, C_N = w, D_N = b, \mathcal{D}) = 4/5$$

Exercise 16.9.12. Fill in the joint probability table for Rule 126.

Solution to Exercise 16.9.12

In working through the results of the last exercise, we saw that Rule 126 was one of the rules that had probability of 1 of outputting both black cells given the causal conditions at the previous time step. Obviously, the numerical assignments to all sixteen cells of the joint probability table under this particular model must have caused this result through the intercession of Bayes's Theorem.

As we have seen many times by now, Bayes's Theorem will take the form,

$$P(B_{N+1} = b \mid A_N = w, B_N = b, C_N = w, \text{Rule 126}) = \frac{P(B_{N+1}, \overline{A}_N, B_N, \overline{C}_N \mid \text{Rule 126})}{P(\overline{A}_N, B_N, \overline{C}_N \mid \text{Rule 126})}$$

$$= \frac{P(B_{N+1}, \overline{A}_N, B_N, \overline{C}_N \mid \text{Rule 126})}{P(B_{N+1}, \overline{A}_N, B_N, \overline{C}_N \mid \text{Rule 126}) + P(\overline{B}_{N+1}, \overline{A}_N, B_N, \overline{C}_N \mid \text{Rule 126})}$$

Construct the joint probability tables in the usual way. The probability in the numerator will be located in cell 6, and the probabilities in the denominator in cell 6 and cell 14. Since the probability of a black cell must be 1, cell 6 must contain a legitimate probability, and cell 14 must contain a 0.

Employ the heuristic of examining the DNF expansion to locate the cells with legitimate probabilities and the cells with 0s. Cells 2 through 7 will contain positive probabilities, and cells 1 and 8 0s. Likewise, cells 9 and 16 will contain legitimate probabilities, and cells 10 through 15 0s. Refer back to Table 16.2 to see that Rule 126 has six terms in its expansion.

The probabilities under the deterministic model of Rule 126 are filled in for the joint probability table as shown below in Figure 16.2.

Figure 16.2: *A joint probability table for Rule 126.*

$$P(B_{N+1} = b \mid A_N = w, B_N = b, C_N = w, \text{Rule } 126) \quad = \quad \frac{\text{Cell } 6}{\text{Cell } 6 + \text{Cell } 14}$$

$$= \quad \frac{1/8}{1/8 + 0}$$

$$= \quad 1$$

The joint probability table for determining C_{N+1} under Rule 126 is exactly the same. The causal factors are different since we have shifted over one cell to the right, but substituting the appropriate cell probabilities results in,

$$P(C_{N+1} = b \mid B_N = b, C_N = w, D_N = b, \text{Rule } 126) \;=\; \frac{\text{Cell 3}}{\text{Cell 3} + \text{Cell 11}}$$

$$= \; \frac{1/8}{1/8 + 0}$$

$$= \; 1$$

Exercise 16.9.13.What is the probability for the other three possibilities that might occur for the two updated cells?

Solution to Exercise 16.9.13

There are four possibilities for the colors of the updated cells B_{N+1} and C_{N+1}. They are 1) (w, w), 2) (w, b), 3) (b, w), and 4) (b, b). Exercise 16.9.11 just found that the probability for the black and black occurrence was $4/5$. Therefore, the other three possibilities must share the probability of $1/5$.

Now, the only rules that output (w, w) are Rules 0, 128, and 192. But these rules were eliminated by the data. The probability for white and white is 0. The only rule that output (b, w) was Rule 85. It also was eliminated by the data so the probability for black and white is also 0. Obviously then, the remaining probability of $1/5$ is allocated to (w, b).

Rule 56 has probability of 1 to output a white cell if the causal factors were white, black, and white. It also has a probability of 1 to output a black cell if the causal factors were black, white, and black. Rule 56 is the only rule to have a probability of 1 for the updated cells to be white and black. Rule 56 was not eliminated by the data, so the averaged prediction for this possibility is $1/5$ as it must be.

Exercise 16.9.14. Discuss the general outline of how you would make an inference about the cell at the previous time step given the output at the current time step.

Solution to Exercise 16.9.14

We are still in the realm of predicting things about abstract ontological systems as represented by an ECA. Here we want to assess a state of knowledge about B_N given B_{N+1}. Once again, we want to illustrate the generality of inferences concerning any statement in the problem space.

Suppose the IP wants to assess its state of knowledge that the previous cell was white given that we know that its updated value is black. Thus, the IP needs to find $P(B_N = w \,|\, B_{N+1} = b)$, or in an even simpler notation, $P(\overline{B} \,|\, Z)$.

Z corresponds to B_{N+1}, A to A_N, B to B_N, and C to C_N. The totality of the model space is indicated by $\mathcal{M}$, and corresponds to the rules running the ECA.

Generically, there are five variables in this problem, A, B, C, Z, and $\mathcal{M}$. Z is the functional output determined by the three arguments A, B, and C. $\mathcal{M}$ stands for the model space. A, B, C, and Z can each assume only two values, while, for the sake of the example, the model space can only take on three values. The dimension of the state space for the possible joint statements is $n = 2 \times 2 \times 2 \times 2 \times 3 = 48$. Thus, a full joint probability table for all five statements would consist of 48 cells.

A direct application of Bayes's Theorem would yield,

$$P(\overline{B} \,|\, Z) = \frac{P(Z\overline{B})}{P(Z)}$$

But these probabilities in the numerator and denominator are sums over other variables. For example, the sum in the numerator explicitly includes terms which are joint statements over relevant variables,

$$P(Z\overline{B}) = P(ZA\overline{B}C\mathcal{M}_1) + P(ZA\overline{B}\,\overline{C}\mathcal{M}_1) + P(Z\overline{A}\,BC\mathcal{M}_1) + P(Z\overline{A}\,\overline{B}\,\overline{C}\mathcal{M}_1) + \cdots$$

$$+ \cdots + P(Z\overline{A}\,\overline{B}\,\overline{C}\mathcal{M}_3)$$

Z and $\overline{B}$ were kept fixed, while A and C were each rotated through their two permissible values. We also have to rotate through the three permissible values for $\mathcal{M}_k$. Thus, there will be a total of 12 terms in the numerator for $P(Z\overline{B})$ for the final summation. There will be four terms involving $\mathcal{M}_2$, as well as the final four terms involving $\mathcal{M}_3$.

In the same manner, the denominator will have these 12 terms showing $\overline{B}$, plus another twelve terms showing B, for a total of 24 terms. Each one of these terms indexes one of the 48 cells in the joint probability table. When the numerical values in these cells are inserted into the terms, the result from Bayes's Theorem will provide the correct quantitative measure on the scale from 0 to 1 of the IP's degree of belief that the statement. "The previous cell was white." is TRUE.

Rather than construct the larger joint probability table to include the model space, we have used the **Product Rule** to separate out $P(\mathcal{M}_k)$. Thus, we can continue to use the smaller 16 cell joint probability table set up for $ZABC$. Of course, the numerical assignments in these 16 cells must change when the model $\mathcal{M}_k$ changes.

Now, the numerator will consist of assignments (four terms in all) from three different 16 cell joint probability tables as conditioned on the kth model. This is multiplied by the probability assigned to each one of these models.

$$P(Z\overline{B}) = P(ZA\overline{B}C \,|\, \mathcal{M}_1)\,P(\mathcal{M}_1) + P(ZA\overline{B}\,\overline{C} \,|\, \mathcal{M}_1)\,P(\mathcal{M}_1) +$$

$$P(Z\overline{A}\,\overline{B}C \,|\, \mathcal{M}_1)\,P(\mathcal{M}_1) + P(Z\overline{A}\,\overline{B}\,\overline{C} \,|\, \mathcal{M}_1)\,P(\mathcal{M}_1) + \cdots$$

$$+ \cdots + P(Z\overline{A}\,\overline{B}\,\overline{C} \,|\, \mathcal{M}_3)\,P(\mathcal{M}_3)$$

Select the numerical assignments to these probabilities from the appropriate cell in each of the 16 cell joint probability tables and insert it into the above expression,

$$P(Z\overline{B}) = (\text{cell 2} \;+\; \text{cell 4} \;+\; \text{cell 6} \;+\; \text{cell 8}) \times 1/3 + \cdots$$

Suppose we pick as models three rules whose joint probability tables we have already filled out in detail. Choose the assignments for Rule 110 appearing in Figure 9.1, the assignments for Rule 30 appearing in Figure 9.2, and the assignments for Rule 126 appearing in Figure 16.2.

The probability in the numerator will then work out to,

$$P(Z\overline{B}) = (1/8 + 1/8 + 1/8 + 0) \times 1/3 +$$

$$(1/8 + 0 + 1/8 + 0) \times 1/3 \;\;+$$

$$(0 + 1/8 + 1/8 + 0) \times 1/3$$

$$= 7/24$$

The probability in the denominator, $P(Z)$, is found in much the same manner by selecting the twenty four numerical assignments appearing in cells one through eight for each of the three rules. Thus, we find that the state of knowledge about the truth of the statement that the cell was colored white at the previous time step, given that it was black when it was updated, is slightly less than $1/2$.

$$P(B_N = w \,|\, B_{N+1} = b) = \frac{7/24}{15/24} = 7/15$$

Appendix A

Introduction to *Mathematica* through Logic Functions

This Appendix is an introduction to the programming language *Mathematica*. Any program referred to in these books is written in this language. *Mathematica* provides first, an unambiguous notation, secondly, the algorithmic steps to a solution, and, finally, a more concise adjunct to the extensive explanations provided in the text. In this Appendix, we use *Mathematica* to help us better understand logic functions as described in Chapter Two.

We will use *Mathematica* to write software to solve problems we can't readily solve by hand. Perhaps, more importantly, we will use it to check that the solutions we did arrive at by laborious hand calculation are, in fact, reaffirmed through a *Mathematica* evaluation.

In addition, it provides us with another notation, a notation that is inherently free of ambiguities. If some program doesn't run as conceptualized in our chosen unambiguous notation (the *Mathematica* syntax), then we know something is amiss.

Wolfram, as the creator of *Mathematica*, naturally enough uses *Mathematica* extensively in his book, *A New Kind of Science*, to provide unambiguous definitions. The present text might be construed as a more elementary introduction to some of those ideas, especially the relation between cellular automata and logic functions.

We begin by fulfilling the promise made at the beginning of the book to provide yet another notation for logic functions. This is the notation employed within *Mathematica* in its role as a general programming language.

In general, all symbolic expressions in *Mathematica* follow this syntax:

$$\texttt{head[arg}_1\texttt{, arg}_2\texttt{, } \cdots \texttt{]}$$

Logic functions of two variables written as symbolic expressions in *Mathematica*

will look, for example, like this,

$$\text{\textbf{And[p, q]}}$$

for what has been written as $A \wedge B$ in the main part of the text. *Mathematica* syntax will be written in a **bold courier font**.

And is the **head** of the symbolic expression with two arguments **p** and **q** which correspond to A and B. The *Mathematica* syntax has been foreshadowed in Chapter Two's discussion of the alternative prefix notation.

Seven of the possible sixteen logic functions for two variables have already been defined in *Mathematica*. The following table, Table A.1, lists these seven built–in functions together with the function numbers, and the symbolic notation as used throughout the main part of the text.

Table A.1: *A list of the seven built–in* Mathematica *logic functions for two variables. The correspondence between the* Mathematica *syntax and the notation used throughout the text is shown as well.*

Row	f_j	*Mathematica* expression	Notation
1	f_2	**Nor[p, q]**	$A \downarrow B$
2	f_5	**And[p, q]**	$A \wedge B$
3	f_8	**Xnor[p, q]**	$A \leftrightarrow B$
4	f_9	**Xor[p, q]**	$A \oplus B$
5	f_{12}	**Nand[p, q]**	$A \uparrow B$
6	f_{13}	**Implies[p, q]**	$A \rightarrow B$
7	f_{15}	**Or[p, q]**	$A \vee B$

So, for example, when we calculated the logic function $A \uparrow B$ by hand for the particular variable settings of $A = T$ and $B = F$, we wrote $f_{12}(T, F)$, and found out that the functional assignment was T by referring to Table 2.5. Double–checking with *Mathematica*, we set **p = True** and **q = False** and then evaluate,

$$\text{\textbf{Nand[True, False]}}$$

As expected, *Mathematica* returns **True**. The *Mathematica* symbols **True** and **False** obviously correspond to T and F.

Since only the above seven functions are pre–defined by *Mathematica*, the other nine logic functions without a corresponding *Mathematica* built–in symbol remain to be defined. The next table, Table A.2, is a listing of some function definitions for the rest of the sixteen logic functions.

Table A.2: Mathematica *definitions for the remaining nine logic functions.*

Row	f_j	*Mathematica* function	Notation
1	f_1	`f1[p_, q_] := False`	$A \perp B$
2	f_3	`f3[p_, q_] := And[Not[p], q]`	$A \star B$
3	f_4	`f4[p_, q_] := And[p, Not[q]]`	$A \diamond B$
4	f_6	`f6[p_, q_] := Not[p]`	$A \vdash B$
5	f_7	`f7[p_, q_] := Not[q]`	$A \dashv B$
6	f_{10}	`f10[p_, q_] := q`	$A \rhd B$
7	f_{11}	`f11[p_, q_] := p`	$A \lhd B$
8	f_{14}	`f14[p_, q_] := Or[p, And[Not[p], Not[q]]]`	$A \leftarrow B$
9	f_{16}	`f16[p_, q_] := True`	$A \top B$

Consider the logic function $A \star B$, which happens to be function $f_3(A, B)$. The full DNF representation for f_3 is $\overline{A} \wedge B$. The syntax for the f_3 function definition in *Mathematica* looks as follows: `f3[p_, q_] := And[Not[p], q]`.

`Not[p]` is *Mathematica* syntax for $\overline{A}$. So this definition is a direct translation of the DNF expression, $\overline{A} \wedge B$, for this logic function.

Here we see that the *Mathematica* syntax is inherently recursive in style. The built–in function `Not[p]` becomes $\textbf{arg}_1$ for the `And` function. But `Not[p]` has its own **head** and one argument.

As another example of the recursive syntactical style, consider function f_{11}. Directly translating the DNF for the $\lhd$ binary operator into *Mathematica*, we have for $(A \wedge B) \vee (A \wedge \overline{B})$, (See Table 2.6, pg. 45),

$$\texttt{f11[p\_, q\_] := Or[And[p, q], And[p, Not[q]]]}$$

One has to pay close attention to the nesting order, and the number and placement of the enclosing brackets in order to arrive at valid *Mathematica* expressions.

The full DNF expression for the $\lhd$ binary operator can be reduced to A by the standard Boolean operations,

$$(A \wedge B) \vee (A \wedge \overline{B}) = A \wedge (B \vee \overline{B}) = A \wedge T = A$$

Therefore, the *Mathematica* expression for f_{11} is simplified to `f11[p_, q_] := p`. The *Mathematica* definitions as shown in Table A.2 are given in the simpler form.

The logic functions evaluate to one of the two symbols `True` or `False`. The function can be directly assigned `True` or `False` as we see for `f1[p_, q_] := False` and `f16[p_, q_] := True` for all variable settings. This, of course, follows from the definition of these constant Boolean functions.

Now it is time to make *Mathematica* play out its role as a means to verify our previous hand calculations. In Chapter Two, one of the preliminary examples was to calculate the answer for the logic expression $(A \star B) \oplus (A \rightarrow B)$ with a setting of $A = T$ and $B = F$. The answer we found back then was F.

What does *Mathematica* say? Translated into a *Mathematica* program, we have,

p = True; q = False; Xor[f3[p, q], Implies[p, q]]

and the answer returned is **False**.

We went to a lot of hard work in the example that followed next where we solved the "weird" looking expression,

$$((A \diamond B) \rhd (A \leftarrow B)) \downarrow ((A \vdash B) \lhd (A \dashv B))$$

for the setting of $A = F$ and $B = T$. Letting *Mathematica* do the work instead, we find that,

p = False; q = True;
Nor[f10[f4[p, q], f14[p, q]], f11[f6[p, q],f7[p, q]]]

does indeed evaluate to **False**.

Notice that in the construction of the *Mathematica* program, we have to be just as careful to pay attention to the nesting order of the functions, and the placement of the brackets as when we did the hand calculation. But still, after having written a correct program, the automatic evaluation provided by *Mathematica* saves a lot of labor.

It is not necessary to specify the settings for the variables **p** and **q**. We can obtain the results for all four variable settings at once for any of the sixteen logic functions by defining another function called **TruthTable**. The *Mathematica* syntax for this very helpful function is,

TruthTable[f_] := Flatten[
                    Outer[f, {True,False}, {True,False}]]

First, we will see it in action and the output it produces. Then, we will discuss its components in more detail.

For a first example, evaluate the functional assignments for all four variable settings for f_1. We already know that this is a constant Boolean function that returns F for any and all variable settings. Thus, if we write **TruthTable[f1]**, *Mathematica* evaluates this as,

{False, False, False, False}.

For a second example, consider the logic function f_8. This is the `EQUAL` operator used extensively in proving tautologies such as,

$$(A \lor B) \leftrightarrow (B \lor A)$$

We simply have to write **`TruthTable[Xnor]`** and *Mathematica* evaluates this correctly as,

`{True, False, False, True}`

Thus, whenever the two variable settings are equal, the function returns **`True`**, and whenever the variable settings are not equal, it returns **`False`**.

We now examine the construction of this **`TruthTable`** function in more detail. The important part is **`Outer[f, {True, False }, {True, False}]`**. The syntax template for **`Outer`** is,

`Outer[`*f, list1, list2, $\cdots$, listn***`]`**.

where *f* is some function and *list1, list2, $\cdots$, listn* represents any number of lists.

In our definition of **`TruthTable`** we have two lists. These two lists are both the same, namely, **`{True, False}`**. The function *f* is applied to each *combination* of the elements comprising each list. Thus, *f* is applied, in turn, to settings of **`p`** and **`q`** of **`{True, True}`**, **`{True, False}`**, **`{False, True}`**, and **`{False, False}`** which is exactly what we want. The function *f* is going to be one of the sixteen logic functions, either one of the built–in *Mathematica* functions like **`Nand`**, or one of the functions which we were forced to define like **`f14`**.

Thus, for **`TruthTable[Xnor]`**, the first evaluation is for **`Xnor[True, True]`** which returns **`True`**, and so on for the other three variable settings. All four results are returned as a list of lists like,

`{{True}, {False}, {False}, {True}}`

The **`Flatten`** function merely strips off this inner set of lists, and returns the answer as shown above.

If the **`TruthTable`** function is understood, it is a short step to defining the new function **`Tautology`**. This function, as it name implies, will automatically evaluate any expression, and determine if it is a tautology. As is our wont, we first observe it working, and then move on to a more detailed examination of its construction.

Above, and in the main text, we started off with the obvious tautology,

$$(A \lor B) \leftrightarrow (B \lor A)$$

Set the left and right sides of the `EQUAL` operator respectively as, **`Or[p, q]`** and **`Or[q, p]`**. Wrap the *Mathematica* implementation of the `EQUAL` operator around this to form **`Xnor[Or[p, q], Or[q, p]]`**.

In order to create these two arguments to **Tautology,** this could be rewritten as **expr[p_, q_] := Xnor[Or[p, q], Or[q, p]]**. This argument is the logic function we wish to test to see if it works out to T for all four variable settings. *Mathematica* evaluates **Tautology[expr]** and returns **True**, thus reaffirming what our hand calculations and intuition originally told us.

Tautology works like this. It takes the output of **TruthTable** and determines whether there is a **False** in the list. If there is, then whatever **TruthTable** worked on cannot be a tautology. **Tautology** uses a function called **FreeQ**. The syntax for **FreeQ** is **FreeQ[***expr*, *form***]**.

For the sake of repetition, this symbolic expression takes the general form of **head[arg$_1$, arg$_2$, $\cdots$]** where **head** is the function **FreeQ** whose two arguments are an expression and a form. It might turn out that the expression is a list like **{True, False, False, True}** and the form is **False**. Is the list free of **False**? No, it is not, so **FreeQ** would return a **False**.

Directly implemented then, **Tautology** is defined as,

```
Tautology[f_] := FreeQ[Flatten[
                Outer[f, {True, False}, {True, False}]], False]
```

where we observe the syntax for **TruthTable** embedded as the *expr* of **FreeQ**.

In the above example testing for a tautology on $(A \vee B) \leftrightarrow (B \vee A)$, **TruthTable** returns the list **{True, True, True, True}**. This expression does not contain **False**, so **FreeQ** returns **True**. We verify that this logic expression is indeed a tautology.

Let's reinforce this by another easy example taken from the main text. The next suspected tautology worked out by hand for one setting of the variables was the expression,

$$((A \rightarrow B) \wedge (B \rightarrow A)) \leftrightarrow (A \leftrightarrow B)$$

Now transition to *Mathematica* syntax for this expression and make it the one argument to **TautologyQ[ ]**. Giving due attention to the proper nesting of the built–in *Mathematica* logic functions, we have,

```
logicExpression[p_, q_] :=
    Xnor[And[Implies[p, q], Implies[q, p]], Xnor[p, q]]
```

Tautology[logicExpression] returns **True** verifying that this expression is indeed a tautology. Running **TruthTable** on **logicExpression** would have returned the list **{True, True, True, True}**.

It is no problem to extend these programs to handle logic functions of three variables. All we need do is add another **{True, False}** to the **Outer** function. This means that all *eight* combinations for the three variables will be examined beginning with **{True, True, True}**, **{True, True, False}**, and ending with **{False, False, False}**.

For example, in section 2.7 we introduced tautologies with three variables and used this example:

$$(A \rightarrow (B \rightarrow C)) \leftrightarrow ((A \wedge B) \rightarrow C)$$

Translate this into *Mathematica* syntax, and form a new argument to yield,

```
logicExpression2[p_, q_, r_] :=
    Xnor[Implies[p, Implies[q, r]], Implies[And[p, q], r]]
```

A **TruthTable3[ ]** function will return a list consisting of eight elements, all **True**, since the above tautology does evaluate to **True** for each one of the eight possible variable settings. Consequently, an analogous function,

```
Tautology3[logicExpression2]
```

will return a **True** since there is no **False** form in the list (expression) returned by a **TruthTable3[ ]** function.

Some further minor observations on *Mathematica* syntax can be mentioned. Symbols like "**=**" for **p = True** and "**:=**" in function definitions make *Mathematica* correspond to the traditional way of thinking and writing about mathematics and computer programs. The symbol "**=**" represents an immediate assignment, while the symbol "**:=**" represents a delayed assignment.

Mathematica offers many alternative notations for expressing the same thing which, in some ways is good, and in other ways not so good. The general syntax can still be used for overall uniformity if so desired. Instead of the conventional **p = True**, one can write **Set[p, True]**. Or, in function definitions, one could write **SetDelayed [***lhs***,** *rhs***]** to maintain the kind of syntactical uniformity as originally shown in **head[arg$_1$, arg$_2$, $\cdots$]**.

As a more complicated example of all this, the general syntax version for the function definition of **TruthTable** would look like,

```
SetDelayed[TruthTable[Pattern[f, Blank[]]],
    Flatten[Outer[f, List[True, False], List[True, False]]]]
```

which, while it does exhibit the virtue of a unifying syntactical structure, calls out for some simplification. These are the syntactical short–cuts like "**:=**" for **SetDelayed**, or "**{$\cdots$ }**" for **List** provided as options in *Mathematica*.

The following two figures and table have been added to this Appendix to help with the translation between Wolfram's presentation of logic functions in *A New Kind of Science* and my presentation. It also helps with translating back and forth between the different numbering systems for the sixteen logic functions. By spending a little bit of time studying Wolfram's numbering system, we can discern an easy visual way of automatically generating the full DNF expansion for any given logic function.

0000	0001	0010	0011
⊥	↓	★	⊢
0	1	2	3
False	Nor	BF[2,2]	BF[3,2]
	$\overline{A}\overline{B}$	$\overline{A}B$	$\overline{A}B \lor \overline{A}\overline{B}$

0100	0101	0110	0111
◇	⊣	⊕	↑
4	5	6	7
BF[4,2]	BF[5,2]	Xor	Nand
$A\overline{B}$	$A\overline{B} \lor \overline{A}B$	$A\overline{B} \lor \overline{A}B$	$A\overline{B} \lor \overline{A}B \lor \overline{A}\overline{B}$

1000	1001	1010	1011
∧	↔	▷	→
8	9	10	11
And	Xnor	BF[10,2]	Implies
AB	$AB \lor \overline{A}\overline{B}$	$AB \lor \overline{A}B$	$AB \lor \overline{A}B \lor \overline{A}\overline{B}$

1100	1101	1110	1111
◁	←	∨	⊤
12	13	14	15
BF[12,2]	BF[13,2]	Or	True
$AB \lor A\overline{B}$	$AB \lor A\overline{B} \lor \overline{A}\overline{B}$	$AB \lor \overline{A}B \lor A\overline{B}$	

Figure A.1: *Wolfram's numbering system for logic functions illustrated with binary operator tables. Pay attention to the fact that the placement of the A and B variables is the opposite of that in the text. Match the ordering of the 1s and 0s in the binary number at the top with the ordering of the black and white cells.*

The first figure, Figure A.1 taking up the whole of the previous page, reproduces one given by Wolfram at the bottom of page 806 in *A New Kind of Science* with some slight amplification on my part. This is the visual way Wolfram chooses to represent the sixteen logic functions with two arguments.

There are sixteen separate little 2×2 tables with each of the four cells filled in as black or white. Each table is an abstract representation of a binary operator table implementing one of the 16 logic functions of two variables. The two rows and two columns are labeled with a black and white square where the black square can be thought of as representing an argument value of TRUE, and the white square as representing an argument value of FALSE for each of the two variables.

The color in a cell is the functional assignment for those particular arguments. Thus, in the very first operator table, $\blacksquare \perp \square = \square$ indicates that the argument to the first variable was $\blacksquare$, the argument to the second variable was $\square$, and that after the binary operation indicated by the symbol $\perp$ was performed, $(\blacksquare \perp \square)$, the resulting functional assignment was $\square$, the white cell in the first row and second column.

I have added the infix operator symbol as previously discussed in Chapter Two in the upper left hand corner of each table as appropriate for the logic function in question. The variable labels A and B are placed differently with A labeling the rows and B the columns. In the text, I usually have it the other way around.

How does Wolfram number the 16 logic functions, and why does he proceed in this manner? The binary number for 0 through 15 is shown at the top of each table. The pattern of 1s and 0s in the binary number as it progresses from 0 to 15 for each logic function encodes whether the cell is black or white with a 1 encoding for black and a 0 for white. The encoding starts at the upper left cell and ends at the lower right cell.

This is explicitly shown in the figure for the two functions **Nor** and **Xor** where the 0s and 1s are placed into each of the four cells. Wherever a 0 appears there must be a white cell, and wherever a 1 appears there must be a black cell, for that logic function.

Moreover, the patterns of 0s and 1s, or the pattern of white and black cells, enables us to immediately write down the full DNF expansion of the logic function in question. The full DNF expansion is written below the operator table for each of the sixteen functions.

For example, the full DNF expansion for the **Xor** logic function is shown below the operator table as $A\overline{B} \vee \overline{A}B$. Wherever a white cell appears that term in the DNF expansion is missing, and wherever a black cell appears that term is present. Thus, the two terms AB and $\overline{A}\,\overline{B}$ are missing in the expansion because they are represented by white cells, and the two terms $A\overline{B}$ and $\overline{A}B$ are present because they are represented by black cells.

Operator Table for the Xnor logic function. It takes two arguments

■ **and** □

where ■ **stands for True and** □ **for False. The infix operator symbol is** ↔

1001

↔ ■ B   ■ B̄

■ A   **1**   0

■ Ā   0   **1**

9

Xnor

AB ∨ ĀB̄

$1001_2 = 9$

The binary representation for the number 9 specifies where black (1) and white (0) will appear in the table.

■ ↔ ■ = ■	`Xnor[True,True]  = True`	$f_8(T,T) = T$
■ ↔ □ = □	`Xnor[True,False] = False`	$f_8(T,F) = F$
□ ↔ ■ = □	`Xnor[False,True] = False`	$f_8(F,T) = F$
□ ↔ □ = ■	`Xnor[False,False]= True`	$f_8(F,F) = T$

Figure A.2: *The 9th logic function according to Wolfram's numbering system which corresponds to my logic function $f_8(A, B)$.*

The second figure, Figure A.2 at the top of the page, provides some more detailed information on the overall picture shown in Figure A.1. For example, consider how Wolfram finds the 9th logic function. The number 9 is represented in binary as 1001 and this indicates that the first cell is black, the second cell is white, the third cell is white, and the fourth cell is black.

Mathematica calls this the **Xnor** function. I have alternatively called it the `EQUAL`, the `BICONDITIONAL`, or the "if and only if" function to correspond to other traditional names for this function in Classical Logic. Evaluating the expression, ■ ↔ □ = □, corresponds to *Mathematica* evaluating **Xnor[True, False]** which returns **False** according to the above table. In Chapter Two, we evaluated an equivalent expression looking like $f_8(T, F) = F$.

The full DNF expansion can be read off from the binary number. The 1s indicate the terms to include and the 0s the terms to exclude. Thus, the canonical, or minterm canonical form, is $AB \vee \overline{A}\,\overline{B}$.

Table A.3 is yet another attempt at providing a map between the various logic expressions and symbols. We have just seen how Wolfram orders the logic functions by binary number. These binary numbers are shown in the second column together with the corresponding name that I used in much of Chapter Two shown in the third column.

My functional notation f_j and operator symbol are shown in the next two columns. The simplified DNF expansion of the function in question is presented next as a change of pace from the full DNF expressions. We match all of this up with the *Mathematica* expression for the logic function in the last column. These are the official names we would employ in all of our *Mathematica* programs.

Table A.3: *Guide for translating between various logic function expressions.*

k	Binary	Name	f_j	Operator	DNF	*Mathematica*
0	0000	FALSE	f_1	$\bot$	F	**False**
1	0001	NOR	f_2	$\downarrow$	$\overline{A}\,\overline{B}$	**Nor**
2	0010	DIFFERENCE	f_3	$\star$	$\overline{A}B$	**BF[2, 2]**
3	0011	NOT A	f_6	$\vdash$	$\overline{A}$	**BF[3, 2]**
4	0100	DIFFERENCE	f_4	$\diamond$	$A\overline{B}$	**BF[4, 2]**
5	0101	NOT B	f_7	$\dashv$	$\overline{B}$	**BF[5, 2]**
6	0110	XOR	f_9	$\oplus$	$A\overline{B} \vee \overline{A}B$	**Xor**
7	0111	NAND	f_{12}	$\uparrow$	$\overline{A} \vee \overline{B}$	**Nand**
8	1000	AND	f_5	$\wedge$	AB	**And**
9	1001	EQUAL	f_8	$\leftrightarrow$	$AB \vee \overline{A}\,\overline{B}$	**Xnor**
10	1010	B	f_{10}	$\triangleright$	B	**BF[10, 2]**
11	1011	IMPLIES	f_{13}	$\rightarrow$	$\overline{A} \vee B$	**Implies**
12	1100	A	f_{11}	$\triangleleft$	A	**BF[12, 2]**
13	1101	IMPLIES	f_{14}	$\leftarrow$	$A \vee \overline{B}$	**BF[13, 2]**
14	1110	OR	f_{15}	$\vee$	$A \vee B$	**Or**
15	1111	TRUE	f_{16}	$\top$	T	**True**

As mentioned before, *Mathematica* provides built–in function names like **Nand** for some of the logic functions. All of these are shown in the appropriate row. What about the logic functions that *Mathematica* does not automatically provide? Earlier, we simply defined new functions for these missing functions.

Mathematica provides another way for us to automatically access these functions without having to explicitly program them as we did before. They are indicated by expressions like **BF[3, 2]** appearing under the *Mathematica* column. This is an abbreviated version for the *Mathematica* expression **BooleanFunction[**k, n**]** which will return the kth Boolean function in n variables. Here, of course, $n = 2$ and k represents the binary number for whatever function that does not possess an already defined name.

For example, suppose we want the 13th logic function for which a ready made function is not available. Set **f13 = BooleanFunction[13, 2]**. Next,

Clear[p, q]; FullForm[BooleanConvert[f13[p, q]]]

returns **Or[p, Not[q]]** which is the unnamed logic function represented by the operator symbol as $A \leftarrow B$. Alternatively, this can be represented by $f_{14}(A, B)$ in the symbology used extensively in Chapter Two when solving problems by hand.

Note that k runs from 0 to 15 instead of from 1 to 16 as shown in the first column of the table. This obviously is to maintain the decimal correspondence to

the binary number.

BooleanConvert[ ] always returns a minimal, or simplified DNF expression, whereas I tend to present the full DNF expansion. **BooleanTable[f13]** returns,

$$\{\texttt{True, True, False, True}\}$$

which corresponds to the pattern of black and white cells for **BF[13, 2]**.

BooleanTable[ ] is a built–in *Mathematica* expression that performs the same task as our previously defined **TruthTable**. **TruthTable[f14]** returns the same list **{True, True, False, True}** for the corresponding logic function f_{14}.

And, finally, as you might have begun to suspect by now, *Mathematica* also provides a ready made tautology testing function called **TautologyQ[ ]**. Our function **Tautology** was defined from scratch so that we could observe the inner workings of such a function. **TautologyQ** will also return the correct value of **True** for the test tautological expressions **logicExpression** and **logicExpression2** defined a few pages ago.

However, there is one important difference. The arguments have to be supplied to any logic expressions for **TautologyQ[ ]** as in,

TautologyQ[logicExpression[p, q]]

TautologyQ[logicExpression2[p, q, r]]

In my custom–designed **Tautology** only the function name had be supplied. The arguments were automatically furnished through the **Outer** function and the lists **{True, False}** that it operated on.

So, in the end, if we have understood the basic mechanics of logic function manipulation from having worked through the calculations in Chapter Two by hand, we can, in the future, shift the burden of all this hard work over to *Mathematica*.

Appendix B

Boolean Functions and *Mathematica*

This Appendix provides some more *Mathematica* syntax for the material presented in Chapter Three. It builds on the introduction given in Appendix A. And, like all the Appendices, this Appendix provides alternative compact presentations for the rather more discursive explanations given in the main part of the text.

Chapter Three was concerned with Boolean functions of three variables in the guise of one–dimensional CA. All told, there are 256 functions with three variables that correspond to Wolfram's 256 rules for CA.

Appendix A introduced the 16 two variable Boolean functions in *Mathematica* notation. Some of these Boolean functions were built–in *Mathematica* functions, while the others were constructed from scratch. There are no automatically built–in *Mathematica* function names for three variable functions.

Simply as an exercise, you may define any of the 256 functions as before. Or, go directly to using **BooleanFunction[**k**, 3]**. Again, each definition will have a counterpart in Wolfram's 256 rules for CA evolution.

Since we now are dealing with three variables, there must be three arguments in the function definition on the left hand side of the *Mathematica* syntax. Make the change to **p, q**, and **r** as the variable names. Analogously to the two variable case, there will be two constant functions that take on the functional assignment of F or T for all variable settings. As before, these are easily defined as,

$$\textbf{rule0[p\_, q\_, r\_] := False}$$

$$\textbf{rule255[p\_, q\_, r\_] := True}$$

These are the counterparts to Wolfram's Rule 0 and Rule 255.

There will be other definitions, just as we have seen before, which are quite simple. For example, Rule 170 is defined as,

```
rule170[p_, q_, r_] := r
```

There are other function definitions which are slightly more complex like Rule 160,

```
rule160[p_, q_, r_] := And[p, r]
```

Finally, there could be a more complicated function definition stemming directly from the full DNF expansion. These will follow a general pattern looking like,

```
fgen[p_, q_, r_] := Or[And[···], And[···], ···]
```

The full DNF expansion for Rule 110 was given in the text as,

$$AB\overline{C} \vee A\overline{B}C \vee \overline{A}BC \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$$

The first term translated into *Mathematica* syntax looks like,

```
And[p, q, Not[r]]
```

The other four terms in the DNF expansion are similarly translated. Putting all five terms into a *Mathematica* function definition, we have,

```
longRule110[p_, q_, r_] := Or[
      And[p, q, Not[r]], And[p, Not[q], r], And[Not[p], q, r],
      And[Not[p], q, Not[r]], And[Not[p], Not[q], r]]
```

So, for example, **longRule110[True, False, True]** evaluates to **True**. Refer back to the third column of Table 3.1 to verify this. That is, in the case of a cellular automaton, a black, white, and black cell at time step N would produce a black cell at time step $N + 1$.

It is not necessary to repeat the **Or** function. In other words, the **Or** function can have as many arguments as needed. From the way *Mathematica* evaluates **Or**, if any of its arguments are **True**, it will return **True**. The same holds for the **And** function where it can be observed that there are three arguments to each of the constituent **And** functions. All arguments must evaluate to **True** for **And** to evaluate to **True**.

In the text, a theorem using the Boolean axioms proved that the full DNF expansion for Rule 110 could be shortened to,

$$B\overline{C} \vee \overline{B}C \vee \overline{A}B$$

Thus we can define a new, shorter version of Rule 110 as,

```
shortRule110[p_, q_, r_] := Or[
            And[q, Not[r]], And[Not[q], r], And[Not[p], q]]
```

In a cellular automaton following the dictates of Rule 110, three white cells should produce another white cell at the next time step. Therefore, our new shorter implementation,

shortRule110[False, False, False]

should evaluate to **False** which it does.

Rather than check just one variable setting at a time, it would be preferable to evaluate the functional assignments for all eight possible combinations of the three variables. In Appendix A, we mentioned that we could easily extend the definition of **TruthTable** to three variables.

```
TruthTable3[f_] := Flatten[
       Outer[f, {True, False}, {True, False}, {True, False}]]
```

TruthTable3[shortRule110] should return with all eight correct functional assignments as listed in Table 3.1. *Mathematica* evaluates this expression as,

{False, True, True, False, True, True, True, False}

which, thankfully, does match up with the functional assignments we previously derived by hand for Rule 110.

Wolfram [20] provides many alternative logic formulas for Rule 110. As an example, in a caption to an illustration of Rule 110 on page 676, he gives this formula,

w1Rule110[p_, q_, r_] := And[Not[And[p, q, r]], Or[q, r]]

Is this **w1Rule110** the same as our **shortRule110**? This is the same as asking whether they produce identical functional assignments for all eight variable settings, which, in turn, is the same as asking whether a tautology exists between these two formulas.

Appendix A mentioned a simple extrapolation that takes us from **TruthTable3** to **Tautology3**. This was done simply for the purpose of showing a low–level *Mathematica* implementation from first principles. However, in all cases where *Mathematica* already has what we want, we will use it instead. So, create a logic expression that is the potential tautology,

```
logicExpression3[p_, q_, r_] := Xnor[
                    shortRule110[p, q, r], w1Rule110[p, q, r]]
```

and use it as the argument in **TautologyQ[logicExpression3[p, q, r]]** which returns the value **True**. Thus, these two formulas are logically equivalent, and return the same functional assignment for any and all variable settings.

On page 884, Wolfram gives a different, and rather involved logic expression for Rule 110,

w2Rule110[p_, q_, r_] := Xor[Xor[And[Not[p], q, r], q], r]

But we can subject this new formula to the same procedure to determine whether it too is logically equivalent to our **shortRule110**. Sure enough, inserting the new argument, **TautologyQ[logicExpression4[p, q, r]]** yields **True** where,

logicExpression4[p_, q_, r_] := Xnor[
shortRule110[p, q, r], w2Rule110[p, q, r]]

This technique helped discover an apparent error in one of Wolfram's formulas for Rule 110. On page 869, Wolfram gives yet another logic expression ostensibly for Rule 110,

w3Rule110[p_, q_, r_] := Xor[Or[p, q], And[p, q, r]]

But, surprisingly, when we run **TautologyQ** on this new expression, a **False** is returned. Upon further examination, **BooleanTable[w3Rule110]** returns,

{False, True, True, True, True, True, False, False}

And so it is clear that such a functional assignment does not match up with Rule 110. Actually, the formula **w3Rule110** is Rule 124 as we found out in Exercise 3.7.4.

As just mentioned, when invoking **TautologyQ** there is no need to go through all the trouble of defining every one of the 256 functions of three variables. We went through that exercise to reinforce the translation of the various notations into *Mathematica* syntax. *Mathematica* automatically provides everything we require without the need for the low–level programming.

We can use **f110 = BooleanFunction[110, 3]**, for example, to define the logic function underlying Rule 110. This is the same built–in *Mathematica* function explained in Appendix A with the only change being the change from $n = 2$ to $n = 3$ arguments.

BooleanTable[f110] returns the list of the eight functional assignments for this Boolean function, so we might as well dispense with our *ad hoc* **TruthTable3**. **FullForm[BooleanConvert[f110[p, q, r]]]** returns the simplified version of the full DNF for Rule 110 as,

Or[And[Not[p], q], And[q, Not[r]], And[Not[q], r]]

thus confirming our hand calculation derived from various Boolean axioms.

There are several other built–in *Mathematica* commands which are quite useful in evaluating logic expressions. The first has the syntax,

Distribute[*expression, overwhat***]**

and it can be used for applying the **Distributivity axiom** when "factoring" and "multiplying" logic expressions. For example,

Distribute[Or[And[p, q], r], And]

will evaluate the expression $(A \wedge B) \vee C$ as $(A \vee C) \wedge (B \vee C)$ analogous to,

$$(x \circ y) \bullet z = (x \bullet z) \circ (y \bullet z)$$

The **LogicalExpand[***expression***]** command will reverse this process. Taking the same example,

FullForm[LogicalExpand[And[Or[p, r], Or[q, r]]]]

will return **Or[r, And[p, q]]**.

LogicalExpand[*expression***]** provides another way of checking for tautologies. Thus, we can use it to check whether the full DNF expansion for Rule 110 is logically equivalent to one of Wolfram's alternative expressions, say, **w2Rule110**.

Both the long rule expression **LogicalExpand[longRule110[p, q, r]]** and Wolfram's expression **LogicalExpand[w2Rule110[p, q, r]]** return (after translation into my notation),

$$\overline{B}C \vee B\overline{C} \vee \overline{A}B$$

verifying first, that the simplification of the DNF as carried out in the text is correct, and secondly, that these two versions of Rule 110 are logically equivalent.

LogicalExpand[*expression***]** will not necessarily return what I have called the full DNF expression, nor does it necessarily return the same answer for different ways of writing the same CA rule. For example, we just saw that **LogicalExpand** returns the same result for **longRule110** and **w2Rule110**.

However, for Wolfram's other expression,

LogicalExpand[w1Rule110[p, q, r]]

returns $\overline{B}C \vee B\overline{C} \vee \overline{A}B \vee \overline{A}C$. A fourth term, $\overline{A}C$, is now present. This is not wrong, but admittedly it is rather peculiar behavior on the part of this command. Exercise 3.7.12 shows the details of why this expression is also correct.

There is a nice concise way of illustrating logical equivalencies between logical expressions for CA rules by constructing the following function. The function will be built up incrementally because *Mathematica* lends itself easily to one command working on the output from a previously applied command.

We know that the output from **BooleanTable[longRule110[p, q, r]]** is a list consisting of **True** and **False**. The command,

Boole[BooleanTable[longRule110[p, q, r]]]

will convert that output to another list of corresponding 1s and 0s. We would like to interpret that list as a binary number. So we apply another command with the first argument the returned output up to this point, and the second argument the base of the number system,

FromDigits[*AlreadyComputedExpression, base***].**

Then,

FromDigits[Boole[BooleanTable[longRule110[p, q, r]]], 2]

will convert that list of 1s and 0s into a binary number. Hopefully, we've retrieved the correct rule number for the expression originally fed into the function.

So construct this function,

checkRuleNumber[expr_] :=
                         FromDigits[Boole[BooleanTable[expr]], 2]

Then feed in as an argument any logic expression that needs to be checked, such as any logic expression for Rule 110, in the form,

checkRuleNumber[longRule110[p, q, r]]

The answer returned is indeed 110.

The tautology between **longRule110** and **w2Rule110** can be double–checked by running **checkRuleNumber[w2Rule110]**. This alternative logic expression also returns a value of 110.

Finally, run this function on Wolfram's suspect logic expression for Rule 110 to confirm our earlier finding. Sure enough,

checkRuleNumber[w3Rule110[p, q, r]]

returns 124, not 110.

Appendix C

Mathematica Programs for Cellular Automata

One of Chapter Three's purposes was to painstakingly detail the inner workings of Wolfram's elementary CA. However, we certainly wouldn't get very far if CA had to be constructed by hand as in those first few illustrative examples.

It is not surprising that the person who created *Mathematica*, and who, at the same time, is one of the world's foremost experts in cellular automata, should provide us with the code to evaluate the evolution of CA. A staggering variety of cellular automata of increasing complexity are covered by this code, but we will examine just what is necessary for evaluating the 256 elementary CA. Just as we did for Boolean functions, we can double–check that our hand calculations of CA evolution are, in fact, verified by Wolfram's code.

We will present this very convenient function at the end of this Appendix. It works like a standard "black box." We feed in the proper ingredients and, in some magical fashion, out comes a visual display of any CA for any number of time steps for any initial starting conditions.

To gain a little more insight into the workings of *Mathematica*, we will examine in detail some of Wolfram's precursor code. These programs were documented in the Technical Notes to Chapter 2 of *A New Kind of Science.* Exploring some of the low–level code behind the work horse function **CellularAutomaton[ ]** permits us a glimpse of how Wolfram tackled cellular automata. This cursory discussion will hopefully advance our understanding of both CA and *Mathematica.*

More advanced examples of cellular automata with more variables and more colors are covered in my Supplemental Exercises accompanying this Volume. As one example, a CA with four colors is explained in order to reproduce Brown's illustration of a "big Boolean Algebra" as broached in the opening Chapter.

The first function discussed is **ElementaryRule**. This function provides a list of the binary numbers for a specified rule. In the text of Chapter Three, we worked this out by hand from fundamental principles. **ElementaryRule** has one argument, the number of the rule we are interested in, and outputs a list consisting of the correct breakdown of the eight 0s and 1s for that decimal number.

It uses a built–in *Mathematica* function called **IntegerDigits**. This function has three arguments as shown in, **IntegerDigits[**n**,** b**,** len**]**, where n is an integer, b converts the integer to base b, and len pads the output with 0s on the left so that the list consists of len elements.

For example, **IntegerDigits[110, 2, 8]** outputs the list,

$$\{0, 1, 1, 0, 1, 1, 1, 0\}$$

This is the breakdown of the integer $n = 110$ into its $len = 8$ constituent digits expressed in base $b = 2$. This output is the same as we obtained from **TruthTable3** for Rule 110 if we match up the 0s with **False** and the 1s with **True**.

ElementaryRule is then simply defined as,

ElementaryRule[num_Integer] := IntegerDigits[num, 2, 8]

so that by evaluating **ElementaryRule[110]** the list above is produced.

Thus, by using this function we can check whether other examples of rules given in the text are correct. Rule 192 was mentioned immediately after Rule 110, and we used this breakdown, among other things, to help write out this rule's DNF. Evaluating **ElementaryRule[192]** yields the list,

$$\{1, 1, 0, 0, 0, 0, 0, 0\}$$

independently verifying our more laborious derivation for this rule.

We will skip lightly over the details for the next two of Wolfram's functions, **CenterList** and **CAStep** because they are not particularly interesting right now. **CenterList** creates the beginning line of cells as a list. For our purposes here, we will just explicitly write out this list as needed.

CAStep is the important function that applies the given rule to the cells at the Nth step, and outputs the revised line of cells for time step $N + 1$. But, again, the details are boring for our present purposes. We will just assume that the function works like a black box to output a list representing the correct colors for the next time step. So, in the example of the CA running according to Rule 110 shown in Figure 3.1, if we evaluate,

CAStep[ElementaryRule[110], {0, 0, 0, 0, 1, 0, 1, 0, 0}]

for the beginning line of cells shown as the list in the second argument to **CAStep**, then the following list is produced,

$$\{0,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 0,\ 0\}$$

representing the second line of cells with the first three cells colored white, followed by four black cells, and finishing with two white cells. This is exactly what the hand generated sketch shows for time step $N + 1$.

Because it illustrates one way *Mathematica* implements iteration, the function **CAEvolveList** is more interesting. Working with the following three arguments, 1) the rule to be applied, 2) the starting configuration for the CA, and 3) the total number of time steps, **CAEvolveList** uses the operator **NestList** for evolving the CA.

Wolfram defines **CAEvolveList[**$arg1,\ arg2,\ arg3$**]** as,

```
CAEvolveList[rule_, init_List, t_Integer] :=
                        NestList[CAStep[rule, #]&, init, t]
```

Applied to our first illustration of a CA operating according to Rule 110 for five time steps, after filling in the arguments it would look like,

```
CAEvolveList[ElementaryRule[110], {0, 0, 0, 0, 1, 0, 1, 0, 0}, 5]
```

The output is a list of six lists with each one of these lists containing the correct 0s and 1s for the color of the nine cells. The first list is the beginning configuration, the next five time steps produce five more lists ending with the last list for the final configuration of the CA.

NestList takes three arguments in the general form of **NestList[**$f,\ x,\ n$**]** where f is some function, x is the argument to the function, and n indicates how many times the function is going to be nested. The output consists of a list with $n+1$ elements with x the first element in the list, $f(x)$ the second element, $f(f(x))$ the third element, and so on, until $\underbrace{f(\cdots f(x)\cdots)}_{n}$ is the final element in the list.

Thus, the first element in the list is the initial list given for the starting line of cells, the second list is the result of **CAStep** working on the initial list and outputting the second line of cells as just shown in the above example at the top of the page.

Since this output of **CAStep** is acceptable input for the second argument of **CAStep**, it can continue on in an iterative fashion. The function **CAStep** is applied n times to each succeeding line of cells. The end result is a list of lists containing the full history of the evolution of the CA over n time steps from the starting configuration to the ending configuration.

When I wanted to check my hand derivation of the CA running according to Rule 110 for 5 time steps as shown in Figure 3.1, I wrote the *Mathematica* expression,

```
CAEvolveList[ElementaryRule[110], {0, 0, 0, 0, 1, 0, 1, 0, 0}, 5]
```

with the result that the list (that is, the history list),

$$\{\{initial\ list\},\ \cdots,\ \{1,\ 1,\ 1,\ 1,\ 0,\ 0,\ 1,\ 0,\ 1\}\}$$

was produced. The final configuration of black and white cells at time step $N + 5$ is shown as the final list.

A subtle point is that `CAStep[rule, #]&` is short–cut syntax for the full form syntax of `Function[x, CAStep[rule, x]]` indicating an unnamed function f that is inserting the correct list as the second argument to `CAStep` as the iteration proceeds. A straightforward substitution of just `CAStep` as the function f would not work because at each iteration `CAStep[rule, init]` would be attempted on the list output at the previous iteration.

Wolfram ends by providing some graphics functions for viewing the list of lists produced by `CAEvolveList`. The evolving CA consists of horizontally stacked lines of cells colored black and white, similar to the sketches in Figures 3.1 and 3.2.

We will finish by briefly discussing the current code provided by *Mathematica* for evaluating the evolution of CA. The function,

```
CellularAutomaton[rule number, starting configuration, time steps]
```

has three self–explanatory arguments. Thus, simply by writing,

```
CellularAutomaton[110, {0, 0, 0, 0, 1, 0, 1, 0, 0}, 5]
```

Mathematica calculates the evolution of an elementary CA according to Rule 110 for five time steps into the future. The starting configuration is given by the list of nine 0s and 1s, with 0 representing a white cell and 1 a black cell.

If you refer back to the sketch of the CA in Figure 3.1, you will see that the starting configuration is four white cells followed by a black cell, a white cell, a black cell, and ending with two white cells. For a visual display of the resulting CA, wrap the `ArrayPlot[ ... ]` function around `CellularAutomaton[ ... ]`.

As already mentioned, more involved examples of `CellularAutomaton[]` are contained in my companion set of worked out exercises to Volume I. An interesting problem for you to ponder is to how to use `CellularAutomaton[]` to create a "toy universe" which runs according to the "Physics" of some unknown logic function.

Appendix D

Proving De Morgan's Axiom with *Mathematica*

In Exercise 5.9.10 of Chapter Five, we presented a rather long and involved proof for **De Morgan's axiom** using logic functions. That is, we proved that the left and right hand sides of the two versions of **De Morgan's axiom** were *logically equivalent*. We devised the function **Tautology** in *Mathematica* for the express purpose of checking for logical equivalency.

We can validate our derivations in that previous exercise with short *Mathematica* expressions. Here, we take advantage of the built–in **TautologyQ[ ]** function as introduced in Appendix A.

Begin by translating from our usual logic notation the left and right hand sides of the first version of **De Morgan's axiom**,

$$\overline{A \vee B} = \overline{A} \wedge \overline{B}$$

into the unambiguous syntax of *Mathematica*.

The left hand side is,

```
Not[Or[p, q]]
```

while the right hand side is,

```
And[Not[p], Not[q]]
```

If we wrap up both these expressions within **Xnor[ ]** to check for logical equivalence between the two expressions, we have,

```
Xnor[Not[Or[p, q]], And[Not[p], Not[q]]]
```

All that's left to do is provide this expression as the argument to **TautologyQ[ ]**,

TautologyQ[Xnor[Not[Or[p, q]], And[Not[p], Not[q]]]]

which returns the answer **True**. *Mathematica* has confirmed our more primitive derivation that the left and right hand sides are equivalent.

The second, dual version discussed illustrating **De Morgan's axiom** was,

$$\overline{A \land B} = \overline{A} \lor \overline{B}$$

This is dealt with in exactly the same manner. The left hand side is,

Not[And[p, q]]

while the right hand side is,

Or[Not[p], Not[q]]

Testing for logical equivalency between these two expressions, we find that,

TautologyQ[Xnor[Not[And[p, q]], Or[Not[p], Not[q]]]]

does indeed return **True**.

The definition of the binary operator NAND is $\overline{A \land B}$. We could substitute the *Mathematica* built–in logic function **Nand[ ]** for the left hand side and re–evaluate,

TautologyQ[Xnor[Nand[p, q], Or[Not[p], Not[q]]]]

Of course, this also evaluates to **True** providing some welcome consistency as we move back and forth between different representations.

Obviously, shifting the burden over to *Mathematica* in order to prove logical equivalency is shorter and more direct than any symbolic proof. For example, we proved in section 5.6 that $P(A \lor B) = P(A \lor \overline{A}B)$. To check this, we have *Mathematica* evaluate,

TautologyQ[Xnor[Or[p, q], Or[p, And[Not[p], q]]]]

TautologyQ, by returning **True**, confirms our rather opaque proof.

Interestingly, and worthy of further investigation, is the converse of what has been demonstrated. A symbolic proof is sometimes shorter than a primitive, brute force method, although ingenuity is then the price one has to pay in order to replace some straight forward mechanical application.

I have a feeling that Wolfram actually has come around to the opinion that "exploration of the computational universe" is, in some sense, to be preferred over mankind's historical reliance on rare mathematical insight and creativity. Those people happily endowed with such aforementioned rare creative genius, say a Roger Penrose for example, would probably disagree.

Here is a final example involving three variables. At the end of Chapter Five, we demonstrated the **Consensus property** as yet another example of some formal symbolic manipulations on abstract probability symbols. There were three variables involved in this proof, and we have presented just two variable examples so far. However, neither **TautologyQ[ ]** nor using **Equivalent[ ]** as a replacement for **Xnor[ ]** has any problem with this wrinkle.

To that end, let's first translate the left hand logic function of the **Consensus property** into *Mathematica* syntax,

$$AB \vee \overline{A}C \vee BC \equiv \texttt{Or[And[p, q], And[Not[p], r], And[q, r]]}$$

Next, translate the logic function on the right hand side,

$$AB \vee \overline{A}C \equiv \texttt{Or[And[p, q], And[Not[p], r]]}$$

Wrap the **Equivalent[ ]** function around the left and right hand sides, and insert it as the argument to **TautologyQ[ ]**,

```
TautologyQ[Equivalent[
                Or[And[p, q], And[Not[p], r], And[q, r]],
                Or[And[p, q], And[Not[p], r]]]]
```

Once again, **TautologyQ[ ]** evaluates this expression to **True**, thus confirming the **Consensus property**.

Appendix E

The *Mathematica* Code Used to Compute the Probability of Future Frequency Counts

The equation below, Equation (15.3) and later as Equation (15.10), was begun in earlier Chapters and exercises, and then finally put together in one place in Exercise 15.7.9. It was used extensively to produce the numerical results appearing in Chapter Fifteen.

$$P(M_1, M_2, \cdots, M_n) = \frac{M!\, \Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})} \times \prod_{i=1}^{n} \frac{\Gamma(M_i + \alpha_i)}{M_i!\, \Gamma(\alpha_i)}$$

It is the formula used to calculate the probability for future frequency counts $M_1, M_2, \cdots, M_n$ appearing in a contingency table with n cells. As is evident in the equation, the probability for these future counts is not conditioned on any further statements. Specifically, they are not conditioned on knowledge of any previous observations, $N_1, N_2, \cdots, N_n$. Thus, this formula applies when no previous data exist. It is called a *prior predictive probability*.

This Appendix contains an annotated presentation of the *Mathematica* code implementing this formula. First, a function,

probabilityOfFutureFrequencyCounts[*arg1, arg2***]**

is defined with lists as the two arguments to the function. The first argument is a list containing the desired frequency counts. The second argument is a list containing the desired α_i parameters in the Dirichlet distribution. The function is defined in the following code.

```
probabilityOfFutureFrequencyCounts[freq_List, alpha_List]:=
        Module[{M, A, ct},
          M = Total[freq];
          A = Total[alpha];
          ct = Factorial[M] Gamma[A] / Gamma[M + A];
          N[ct Apply[Times, (Gamma[freq + alpha]) /
                              (Factorial[freq] Gamma[alpha])]]
                                                    (* close Module *) ]
```

A function call with typical arguments to the function might look like,

```
probabilityOfFutureFrequencyCounts[{4, 4, 4, 4}, {1, 1, 1, 1}]
```

which evaluates to a value of 1.03×10^{-3} as shown in Table 15.1.

Here we are computing the probability of the future frequency count consisting of four kangaroos in each of the four cells indexing a beer–hand preference. The total number of kangaroos in the contingency table is 16. The second argument implements a uniform distribution over model space by prescribing that all four α_i parameters equal 1. Every single model making a legitimate numerical assignment to the probability of the four joint statements in the state space is given equal credence. No one model is favored over another.

The code for the function definition is surprisingly short and transparent. The left hand side is the standard short cut for a delayed *Mathematica* function definition **SetDelayed[***lhs, rhs***]** where the *lhs* is **f[arg$_1$, arg$_2$]** and the *rhs* is everything beginning with **Module [$\cdots$]**.

On the right hand side, everything is wrapped in a **Module [$\cdots$]** because three local variables **M, A,** and **ct** are defined. These local variables stand for the sum of the frequency counts, the sum of the α_i parameters, and the constant term in the predictive equation.

Thus, the **Total[***list***]** function,

```
M = Total[freq];
A = Total[alpha];
```

tallies up the sums of the frequency counts and the alpha parameters. Both the frequency counts and the alpha parameters are contained in lists as illustrated above when we discussed them as the arguments to the function.

As would be expected, *Mathematica* has built–in factorial and Gamma functions for computing terms like $M!$ and $\Gamma(\mathcal{A})$. Thus the constant term C in the predictive equation,

$$C = \frac{M! \, \Gamma(\mathcal{A})}{\Gamma(M + \mathcal{A})}$$

is written in *Mathematica* code as,

```
ct = Factorial[M] Gamma[A] / Gamma[M + A];
```

Multiplication in *Mathematica* is implicit by writing the two terms in the numerator separated by a space. When the **FullForm[** $\cdots$ **]** expression is evaluated,

ct = Times[Factorial[M], Gamma[A], Power[Gamma[Plus[A, M]], -1]]

the syntactical shortcuts for addition, multiplication, and division are evident.

The only somewhat tricky part of the code occurs in the final line of code. Here, *Mathematica*'s **Apply[***f, expr***]** construct is used to carry out the multiplications in all of the terms beyond the constant term. This is the,

$$\prod_{i=1}^{n} \frac{\Gamma(M_i + \alpha_i)}{M_i! \, \Gamma(\alpha_i)}$$

part of the equation.

Since both **freq** and **alpha** are lists, all the computation taking place in the remaining product terms beyond the constant term are also lists. Thus, they have the head **List[** $\cdots$ **]** attached to them. The function **Apply[** $\cdots$ **]** says to take the first argument *f*, here **Times**, and replace the head of the second argument *expr*, that is, **List[** $\cdots$ **]**, with **Times**. In effect, we multiply all the elements in the list which is exactly what we want to do.

The **N[***expr***]** function is employed to convert what potentially might still be expressed as a fraction into a numeric value. Interestingly, if **N[** $\cdots$ **]** is omitted, the function returns a value of 1/969, a very familiar number. This final expression is the actual output from the function, and provides us with the numerical value of the probability for the specified frequency counts.

The semicolons appearing at the end of lines is the way *Mathematica* suppresses the output of any computation. The only output we want to appear is the final line.

Because *Mathematica* predisposes one to write highly nested code, one must constantly be cognizant of closing off all of the nested function calls. For example, the final right bracket (**]**) is shown on its own separate line to indicate that it closes off the **Module[** $\cdots$ **]** function. The previous two right brackets close off, respectively, the **Apply[** $\cdots$ **]** function and the **N[** $\cdots$ **]** function.

A comment is in order concerning the expressions **freq_List** and **alpha_List** appearing as the arguments to the function. Written in this way, the arguments appear in an acceptable alternative shortened form. Consider a function of one argument written with the typical syntactical short–cuts as,

```
afunction[a_List] := Total[a]
```

A function call would look something like **afunction[{1, 2, 3}]** and evaluates to 6.

Alternatively, this function can be constructed as a longer general *Mathematica* symbolic expression using the nested **head[arg$_1$, arg$_2$, $\cdots$]** template, where **arg$_2$** might itself be another **head[arg$_1$, arg$_2$, $\cdots$]** expression, and so on to whatever level of nesting might be required.

Thus, in writing,

SetDelayed[afunction[Pattern[a, Blank[List]]], Total[a]]

we observe that **arg$_2$** for the head **SetDelayed[ ]** is **Total[a]**, and **arg$_1$** is **afunction[Pattern[a, Blank[List]]]** illustrating further nesting patterns.

Since **Blank** now has the argument **List**, the argument is constrained to be a list. No syntactic short–cuts are employed in this representation. The expression **afunction[a]** would return unevaluated in exactly the same form in which it was submitted, but **afunction[{b, c, d}]** evaluates to **b + c + d**.

Appendix F

Glossary

Here is a review of some symbols and important terms in probability and inferencing as used over the course of this Volume.

Statements. $A, B, C, \ldots$ refer to statements which must be either TRUE or FALSE. A statement is always indicated by enclosing it in quotation marks and ending it with a period.

$A \equiv$ "The coin shows HEADS."
$B \equiv$ "The die shows a FIVE spot."
$C \equiv$ "The kangaroo drinks Foster's beer."

False statements. The *overbar* symbol placed over $A, B, C, \ldots$ indicates that the statement is FALSE.

$\overline{A} \equiv$ "The coin shows HEADS." is FALSE.
$\overline{B} \equiv$ "The die shows a FIVE spot." is FALSE.
$\overline{C} \equiv$ "The kangaroo drinks Foster's beer." is FALSE.

Observed values. $(A = a_i), (B = b_j), \ldots$ refer to a measured or observed value in the statement. If A can take on only two values, then $(A = a_1)$ or $(A = a_2)$. An alternative notation as used in logic was $(A = T)$ or $(A = F)$.

$A \equiv (A = T) \equiv (A = a_1)$ "The coin shows HEADS."
$\overline{A} \equiv (A = F) \equiv (A = a_2)$ "The coin shows HEADS." is FALSE, or equivalently, "The coin shows TAILS."

Probability operator. The probability operator is wrapped around a statement, $P(A = a_i)$, to indicate an IP's degree of belief that the statement $(A = a_i)$ is TRUE. This is a symbolic probability expression with no number involved.

$P(B = b_5)$ is an abstract and symbolic representation for an IP's degree of belief that the statement, "The die shows a FIVE spot." is TRUE.

Numerical probability 1. Numerically, a probability must be a real number (as opposed to a complex number) between 0 and 1 inclusive.

$P(C = c_2) = 1/2$ indicates that the IP's degree of belief that the statement, "The kangaroo drinks Corona beer." is TRUE, is mid–way on the scale from 0 to 1.

Numerical probability 2. The probability assignment of 0 or 1, $P(A) = 0$ or $P(A) = 1$, must also be included to indicate certainty that A is FALSE or certainty that A is TRUE.

$P(A) \equiv P(A = a_1) = 0$ indicates that the IP believes that the statement, "The coin shows HEADS." is certainly FALSE.
$P(A) \equiv P(A = a_1) = 1$ indicates that the IP believes that the statement, "The coin shows HEADS." is certainly TRUE.
$P(\overline{A}) = 0 \equiv P(A = a_2) = 0$ "The coin shows TAILS." is certainly FALSE is equivalent to "The coin shows HEADS." is certainly TRUE, or $P(A) = 1$.
$P(\overline{A}) = 1 \equiv P(A = a_2) = 1$ "The coin shows TAILS." is certainly TRUE is equivalent to "The coin shows HEADS." is certainly FALSE, or $P(A) = 0$.

Numerical probability 3. Reference to a probability as a numerical assignment must include conditioning on some model $\mathcal{M}_k$. Conditioning on a model is written as $P(A = a_i \,|\, \mathcal{M}_k)$. The numerical value is $Q_i = P(A = a_i \,|\, \mathcal{M}_k)$.

$P(B = b_3 \,|\, \mathcal{M}_1) = 1/6$ indicates the IP's degree of belief on the scale from 0 to 1 that the statement, "The die shows a THREE spot." is TRUE. This degree of belief is conditioned on assuming model $\mathcal{M}_1$ is TRUE. Model $\mathcal{M}_1$ assigns what it considers as the "correct" numerical values to the six probabilities for B.

Model statement. $\mathcal{M}_k$ is also a statement.

The statement is something like, "The correct, indisputable, and even irrevocable assignment of a legitimate numerical probability to the statement $(A = a_i)$ is Q_i." Such an assignment is based on unique *information* about $(A = a_i)$.

Probability for a model. Expresses the IP's implicit degree of belief that the specified model assigning the "correct" numerical values as probabilities to the statements in the state space is TRUE.

If model $\mathcal{M}_k$ were to assign probability 3/4 to $(A = a_1)$ and 1/4 to $(A = a_2)$, then $P(\mathcal{M}_k) = 1/5$ might be the IP's degree of belief that this assignment is TRUE.

Numerical probability 4. A probability is a measure of the degree of belief about the truth of one statement conditioned on the assumed truth of another given statement.

$P(C = c_1 \mid \mathcal{M}_2) = Q_1 = 3/4$ represents the IP's degree of belief about the truth of "The kangaroo drinks Foster's." dependent on assuming that a second model has made a statement about "correct" numerical assignments for the two values of C.

Conditional probability 1. Same as above.

The conditional probability $P(A = a_1 \mid B = b_2, \mathcal{M}_k)$ is the IP's degree of belief that statement $(A = a_1)$ is TRUE given that statement $(B = b_2)$ is assumed TRUE and that model $\mathcal{M}_k$ that made the numerical assignments to A and B is also assumed TRUE.

Joint statement. Two or more statements joined by AND.

$(A = a_i)$ AND $(B = b_j)$ AND $\ldots$
A probability for a joint statement is written as $P(A = a_i, B = b_j, \cdots \mid \mathcal{M}_k)$.

Numerical probability 5. Any numerical probability for a statement A whose observed value is a_i, written as $P(A = a_i)$ without any explicit reference to a model statement, must be understood as,

$$P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i \mid \mathcal{M}_k) \, P(\mathcal{M}_k)$$

where the symbol $\mathcal{M}$ refers to the total number of models.

The degree of belief that "The coin shows HEADS." is TRUE is $1/2$ when the degree of belief is averaged over the models assigning all possible values from 0 through 1 to this statement.

Dimension of state space. The integer n refers to the total number of joint statements in the state space.

If statement A can assume two measured values $(A = a_1)$ and $(A = a_2)$, while statement B can assume three measured values, $(B = b_1)$, $(B = b_2)$, and $(B = b_3)$, there are in total $n = 6$ possible joint statements $(A = a_i, B = b_j)$.

Frequency count 1. The integers $N_1, N_2, \cdots, N_n$ refer to *past* frequency counts.

N_1 refers to the number of times that the first joint statement in the state space was true, N_2 the number of times that the second joint statement in the state space was true, and N_n refers to the number of times the nth, and final, joint statement in the state space was true. $N_1, N_2, \cdots, N_n$ have actually already happened, so there is no uncertainty about their occurrence.

Data. The $N_1, N_2, \cdots, N_n$ past frequency counts. The symbol $\mathcal{D}$ refers to the data.

N. The sum $N = N_1 + N_2 + \cdots + N_n$ represents the total number of past frequency counts, or the total number of data points.

Frequency count 2. The integers $M_1, M_2, \cdots, M_n$ refer to *future* frequency counts.

M_1 refers to the number of times that the first joint statement in the state space will be true, M_2 the number of times that the second joint statement in the state space will be true, and M_n to the number of times the nth, and final, joint statement in the state space will be true. $M_1, M_2, \cdots, M_n$ have NOT actually already happened so there IS uncertainty about their occurrence.

M. The sum $M = M_1 + M_2 + \cdots + M_n$ represents the total number of future frequency counts. A probability operator may be wrapped around the joint statement $P(M_1, M_2, \cdots, M_n)$ to indicate that there are no previous data. On the other hand, $P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n)$ indicates a state of knowledge conditioned on N data points.

A trial. A measurement or observation is made on a statement in the state space. The temporal or other ordering is affixed as a subscript.

$(A_1 = a_2)$ "TAILS was observed on the first trial of the toss of the coin."
$(B_2 = b_4)$ "A FOUR spot was observed on the second roll of the die.", or,
　　　　"A FOUR spot was observed on the roll of the *red* die."
when two dice, a green die and a red die, were rolled at the same time.

The next trial. After N trials have taken place, and N observations have been made, a probability required for the next trial is written $P(A_{N+1})$. Thus, no measurement or observation has yet been made for the $(N+1)^{st}$ trial.

The data from $N = 30$ students concerning test scores and graduation have been recorded. The graduation status for the 31st student is not yet known. The proper observation determining the truth of this statement constitutes the next trial.

The next M trials. After N trials have already taken place, and N observations have already been made, the probability for M future trials is written as $P(A_{N+1}, A_{N+2}, \cdots, A_{N+M})$.

The data from $N = 30$ students concerning test scores and graduation have been recorded. The graduation status for the next $M = 10$ students is not yet known and they constitute the next M trials. A probability for the 31st student, 32nd student, $\cdots$, and the 40th student is required.

The first trial. When $N = 0$ and $M = 1$, $P(B_{N+M} = b_6) \equiv P(B_1 = b_6)$.

If a die has not been rolled in the past, the probability for the next M rolls to show the SIX spot, where $M = 1$, is the probability for the very first trial to show a SIX spot, $P(B_1 = b_6)$. If $M = 2$, the probability for the green die to show a THREE spot and the probability for the red die to show a FIVE spot is $P(B_1 = b_3, B_2 = b_5)$ when the two dice are rolled together.

Numerical probability 6. The numerical probability for the next trial after N data points is,

$$P(A_{N+1} = a_i \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} [\, P(A_{N+1} = a_i \,|\, \mathcal{D}, \mathcal{M}_k) \times P(\mathcal{M}_k \,|\, \mathcal{D}) \,]$$

After the results from $N = 100$ tosses of a coin have been recorded, a probability that the next toss is TAILS is an average of the probability for TAILS as made by all $\mathcal{M}$ models. The weighting given to each model's prediction for TAILS is dependent on the data's support for that model. *E.g.*, $P(A_{101} = a_2 \,|\, \{A_1, A_2, \cdots, A_{100}\}) = 5/8$.

Numerical probability 7. The numerical probability for the next trial after N data points, and conditioned on the truth of the kth model, is,

$$P(A_{N+1} = a_i \,|\, \mathcal{D}, \mathcal{M}_k) \equiv P(A_{N+1} = a_i \,|\, \mathcal{M}_k) = Q_i$$

The assignment of a legitimate numerical value to a probability depends only on the model, and is independent of any past data. After the results from $N = 100$ tosses of a coin have been recorded, a probability that the next toss is TAILS conditioned on assuming true model 27's assignment of numerical values is, say, $3/8$.

$$P(A_{101} = a_2 \,|\, \{A_1, A_2, \cdots, A_{100}\}, \mathcal{M}_{27}) \equiv P(A_{101} = a_2 \,|\, \mathcal{M}_{27}) = Q_2 = 3/8$$

Numerical probability 8. The numerical probability for the next trial after N data points depends on how the probability for the models has been changed because of the data.

If model $\mathcal{M}_1$ asserts that $Q_1 = 1$ and TAILS shows up on the 100th toss of the coin, then $P(\mathcal{M}_1 \,|\, \mathcal{D}) = 0$. Model 1's prediction of a probability of 1 for HEADS on the next toss is not counted in the averaging process.

Contingency table. A table consisting of n cells showing either the N past data, or the M future frequency counts. A count of N_1 is placed in cell 1, a count of N_2 is placed into cell 2, $\cdots$, and a count of N_n is placed into cell n.

The state space for, say, the kangaroo scenario consists of the four joint statements $(A = a_i, B = b_j)$. The dimension of the state space is $n = 4$. Every contingency table for this scenario consists of $n = 4$ cells. A possible contingency table contains these future frequency counts, $M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1$. The first cell $(A = a_1, B = b_1)$ contains a future frequency count of 9 kangaroos, and the nth cell $(A = a_2, B = b_2)$ contains a future frequency count of one kangaroo. The total number of future kangaroos placed into the n cells of the contingency table is $M = M_1 + M_2 + \cdots + M_n = 16$.

Joint probability table. The same kind of table as the contingency table, but instead of holding frequency counts, the cells contain assigned probabilities.

Cell 1 of the joint probability table for the kangaroo scenario contains the probability for a joint statement given that some model is assumed true. Q_1 is the probability for the joint statement when model $\mathcal{M}_k$ has made the assignments,

$$P(A = a_1, B = b_1 \,|\, \mathcal{M}_k) = Q_1$$

Likewise, when the dimension of the state space is $n = 4$, cell n contains,

$$P(A = a_2, B = b_2 \,|\, \mathcal{M}_k) = Q_n = Q_4$$

Joint probability. The probability in each of the n cells comprising the joint probability table.

$P(A = a_1, B = b_2 \,|\, \mathcal{M}_k) = Q_3$ is a joint probability in the third cell of some joint probability table. For the kangaroo scenario, Q_3 represents a degree of belief, $P(\overline{R}F)$, that the joint statement, "The kangaroo drinks Foster's beer and uses its left hand." is TRUE under the information in model $\mathcal{M}_k$.

Marginal probability. The probability appearing at the margins of the joint probability table. It is the sum of the individual joint probabilities in the cells to the left of, or above where the marginal probability appears.

$P(A = a_1 \,|\, \mathcal{M}_k) = P(A = a_1, B = b_1 \,|\, \mathcal{M}_k) + P(A = a_1, B = b_2 \,|\, \mathcal{M}_k)$ is a marginal probability appearing at the bottom of a 2×2 joint probability table under the column labeled as A, and is the sum of the two cells above it. In general,

$$P(A = a_i \,|\, \mathcal{M}_k) = \sum_{j=1}^{n_B} P(A = a_i, B = b_j \,|\, \mathcal{M}_k)$$

where n_B refers to the total number of possible measurements on B, *i.e.*, the number of rows for B in the joint probability table.

Conditional probability 2. A conditional probability may also be calculated via Bayes's Theorem as the ratio of a joint probability over a marginal probability.

$$P(A = a_i \,|\, B = b_j, \mathcal{M}_k) = \frac{P(A = a_i, B = b_j \,|\, \mathcal{M}_k)}{P(B = b_j \,|\, \mathcal{M}_k)}$$

Since $P(B = b_j \,|\, \mathcal{M}_k)$ is a marginal probability, it gets expanded into,

$$P(B = b_j \,|\, \mathcal{M}_k) = \sum_{i=1}^{n_A} P(A = a_i, B = b_j \,|\, \mathcal{M}_k)$$

where n_A refers to the total number of possible measurements on A, *i.e.*, the number of columns for A in the joint probability table.

State space. The set of all n statements (possibly joint) defining what can be measured or observed during any trial.

> The four joint statements, (1) "A kangaroo drinks Foster's beer and is right–handed." (2) "A kangaroo drinks Corona beer and is right–handed."(3) "A kangaroo drinks Foster's beer and is left–handed."(4) "A kangaroo drinks Corona beer and is left–handed." constitute the state space for the kangaroo scenario. The measurement of beer preference and hand preference are the only two observations that will be made on an individual kangaroo. The observation on any one individual distinguishable kangaroo constitutes a trial.

n. The dimension of the state space. The total number of joint statements of which only one might be observed as TRUE during a trial.

> The dimension of the arbitrarily defined state space for the die tossing scenario is $n = 6$. Either a ONE face, or a TWO face, $\cdots$, or a SIX face, and only one of these, can be observed as TRUE on any trial.

Sample space. The expression n^M indicates the number of elementary points in the sample space.

> Let the dimension of the state space describing a coin toss be $n = 2$. Suppose that the IP is concerned with a total of $M = 4$ *future* tosses of the coin. There are $2^4 = 16$ elementary points in the sample space of a coin that will be tossed four times.

Elementary point. The most detailed statement that can be made about events in the sample space.

> One of the 16 elementary points in the above sample space is, "HEADS appears on the first toss, TAILS appears on the second toss, TAILS appears on the third toss, and HEADS appears on the fourth and final toss." No event in the sample space can be more detailed than this.

Event. An aggregation of some number of elementary points existing in the sample space.

> The event, "Two HEADS and two TAILS in four tosses of the coin." represents an aggregation of six elementary points in the sample space. One of these six elementary points was given in the previous entry.

State of knowledge. Synonymous with the distribution of numerical probabilities over whatever space is under consideration.

> The IP's state of knowledge about the number of HEADS to appear in the next four tosses of the coin when the IP is completely uninformed is given a quantitative status with a probability of 1/5 for no HEADS, 1/5 for one HEADS, 1/5 for two HEADS, 1/5 for three HEADS, and 1/5 for four HEADS.

Bibliography

[1] Aitchison, J. and Dunsmore, I. R. *Statistical Prediction Analysis*, Cambridge University Press, Cambridge, UK, 1975.

[2] Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*, John Wiley & Sons, Ltd., Chichester, England, 1994.

[3] Brown, Frank Markham. *Boolean Reasoning: The Logic of Boolean Equations*, Second Edition. Dover Publications, Mineola, NY, 2003.

[4] Feller, William. *An Introduction to Probability Theory and Its Applications, Volume I*, 3rd Edition, Revised printing. *Volume II*, Second Edition, John Wiley & Sons, New York, NY, (Vol I) 1968 (Vol II) 1971.

[5] Frieden, B. Roy. Unified theory for estimating frequency–of–occurrence laws and optical objects. *J. Opt. Soc. Am.*, Vol. 73, No. 7, 1983.

[6] Garrett, Anthony J. M. Probability Synthesis: How to Express Probabilities in Terms of Each Other. *Proceedings of the 17th International Workshop on Maximum Entropy and Bayesian Methods.*, Boise, ID, 1997, pp. 115–120, Kluwer Academic Publishers, Dordrecht, 1998.

[7] Geisser, Seymour. *Predictive Inference: An Introduction.* Chapman & Hall, New York, NY, 1993.

[8] Gull, S. F. and Skilling, J. *The Maximum Entropy Method*, in Indirect Imaging, ed. by J. A. Roberts, Cambridge University Press, 1984.

[9] Hald, Anders. *A History of Mathematical Statistics from 1750 to 1930.* Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, 1998.

[10] Hofstadter, Douglas R. *Gödel, Escher, Bach: an Eternal Golden Braid.* Basic Books, New York, 1979.

[11] Hohn, Franz E. *Applied Boolean Algebra*, Second Edition, The MacMillan Company, New York, NY, 1966.

[12] Jaynes, Edwin T. *Probability Theory: The Logic of Science*, ed. by G. Larry Bretthorst, Cambridge University Press, New York, NY, 2003.

[13] Jaynes, Edwin T. Monkeys, Kangaroos, and N. In *Maximum Entropy and Bayesian Methods in Applied Statistics.* ed. by J. H. Justice, pp. 27–58, Cambridge University Press, 1986.

[14] Jeffreys, Harold. *Theory of Probability*, Third Edition, Oxford University Press, 1961.

[15] Kass, R. E. and Wasserman, L. Formal rules for selecting prior distributions: A review and annotated bibliography. *J. Amer. Statist. Assoc.*, 91, pp. 1343–1370, 1996.

[16] Mendelson, Elliott. *Boolean Algebra and Switching Circuits*, Schaum's Outline Series in Mathematics, McGraw–Hill, New York, NY, 1970.

[17] Schneeweiss, Winfred G. *Boolean Functions with Engineering Applications and Computer Programs*, Springer–Verlag, Berlin, Germany, 1989.

[18] Smullyan, Raymond. *Forever Undecided*, Alfred A. Knopf, New York, NY, 1987.

[19] Tipler, Frank J. *The Physics of Immortality: Modern Cosmology, God and the Resurrection of the Dead*, Doubleday, New York, NY, 1994.

[20] Wolfram, Stephen. *A New Kind of Science*, Wolfram Media, Inc., Champaign, IL, 2002.

Index