INFORMATION PROCESSING

Supplemental Exercises

for

Volume I

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

David J. Blower

© *David J. Blower, Third Millennium Inferencing,*
Pensacola, Florida, February 2018.

DEDICATED TO THE MEMORY
OF MY FAITHFUL WALKING COMPANION

LILY

Preface

During my development of core concepts for probability and inferencing in Volume I, far too many interesting and solvable problems presented themselves than could possibly be included in the space available. Rather than simply abandoning them, I decided that a companion book would have to be written containing many more fully solved exercises to supplement the ones already appearing in Volume I.

As emphasized in Volume I's *Apologia*, I am a firm believer in thinking that the very best path to learning for us normal mortals has to be in the steady consumption of a vast array of solved numerical examples. My basic inclination is to solve *easy* problems first.

It is only then, in thinking through and solving progressively more complicated exercises, that any kind of global awareness eventually sinks in. When we run into the inevitable barriers to solving these increasingly more difficult exercises, we then know the boundaries of our current imperfect knowledge. The reward is perhaps a clue pointing to what new knowledge must be created to expand these boundaries.

Besides that, I think one is entitled to some bracing positive reinforcement from time to time. For me, and I'm sure for many others, that pleasant experience occurs when I successfully solve a problem where the resulting numbers possess the subjective impression that they must be correct. If I can perceive any abstract theoretical development through the lens of a series of computations I feel rewarded. Nevertheless, always keep in mind the saying that the purpose of computation is not in the numbers themselves, but in any *insight* that the numbers provide.

It is my strongly held belief that creating and solving problems when generated on our imperfect understanding of core concepts is the best pedagogical device in existence. I wish that mathematics were taught more in this manner. Let each student make up his or her own set of problems as simple or complex as may be, but always starting by crafting an exercise that the student already supposedly knows how to solve.

Thus, the problems solved here in these Supplemental Exercises are somewhat harder or more involved than those that have already appeared in Volume I's set of exercises. I also indulge in the luxury of taking as much space as I desire to fully explicate some core concepts that could only be touched upon lightly in Volume I.

For example, I devote a full eight pages and four separate exercises to what I consider a thorough dissection of Feller's analysis of *Probability of Accidents During the Week*. This effort supplements my presentation of Feller's counting formulas necessary for the solution of classical occupancy problems in Chapter Thirteen. I am sure that there must be many like me who have labored over Feller's typically condensed explanation of how to assign probabilities to the number of accidents that happen during the course of a week. It took me a very long time before I fully understood how all of his numbers made sense.

Thus, these supplemental exercises might be charitably viewed as my first step in setting up slightly more complicated inferencing problems. As I said, the goal is to expand the boundaries. Nonetheless, I have tried to build on the original exercises so that introducing new complications did not open up any novel mysterious conceptual gaps. I have attempted to make each foot hold and hand hold climbing up the cliff as accessible as possible.

But this task is impossible. The creation of new knowledge cannot in general be reduced to small, smooth, easily traversed incremental steps. It is more of a slow evolution with long periods of trivial extensions punctuated hopefully every now and then with *conjectures* that are more a leap of faith. These conjectures must then be subjected to intense *criticism* to winnow the wheat from the chaff.

As an example of that kind of conjectural effort on my part, I spend some time on examining the analogy between de Finetti's Representation Theorem and a commonly used probability manipulation rule. This is the rule that expresses the probability for a statement as a weighted average over all of the models lending a numerical assignment to that statement.

By way of the most obvious linkage back to Volume I, the Chapter Titles for these Supplemental Exercises exactly match those of Volume I. As you might have expected, I have made an honest atempt to restrict the exercises to the topics of each Chapter. Nonetheless, I do take the liberty from time to time of referring to developments that actually appear later on than in the current Chapter, as well as to a few concepts that only make their appearance in Volumes II or III.

David John Blower
Pensacola, Florida
February 2018

Contents

Preface **i**

1 Boolean Algebra **1**

1.1 Notation 1

1.2 Three Variable Boolean Functions 5

1.3 Boole's Expansion Theorem for Three Variable Functions 8

2 Logic Functions **11**

2.1 Jaynes's Plausible Reasoning 11

2.2 Another Logic Puzzle 16

2.3 Revisiting the First Logic Puzzle 25

3 Cellular Automata **29**

3.1 ECA and the Second Logic Puzzle 29

3.2 Rule 59 and the CNF 32

3.3 Three Color CA 40

4 Analogies Between Formal Manipulations **51**

4.1 Ordering Relations in Boolean Algebra 51

4.2 Analogies with Probability Assignments 55

5 Fundamental Rules of Probability **59**

5.1 Boolean Operations and Probability 60

5.2 Conceptual Foundations 64

5.3 Solving Jaynes's Exercise 2.1 73

5.4 Solving Jaynes's Exercise 2.2 75

6 Bayes's Theorem **77**

6.1 Status of Bayes's Theorem 77

6.2 Alternative Versions of Bayes's Theorem 81

6.3 Logic Functions as Models 87

6.4 Exercise 6.6.18 of Volume I 98

6.5 Jaynes and Sampling without Replacement 102

6.6 Non–binary Variables 121

7 Generalizing Logic with Probability **125**

7.1 Revisiting the Logic Puzzle from Chapter Two 125

7.2 *Mathematica* and Some Classical Syllogisms 128

7.3 Revisiting the Life on Mars Scenario 131

7.4 Where are the Data? 137

8 Deterministic Cellular Automata **139**

8.1 Enforcing Determinism 139

8.2 More Complex Deterministic CA 147

9 Probabilistic Cellular Automata **155**

9.1 Probabilistic CA: An Oxymoron 155

9.2 Correct Involvement of Probability 158

9.3 Relevant Cell Colors Are Unknown 160

9.4 Information about Rules 162

10 Logic Puzzles **169**

10.1 The Halloween Party Logic Puzzle 169

10.2 Clausal Form Logic 175

11 Formal Rules for Prediction **183**

11.1 Coherent Arguments 183

11.2 Coherency with Myself? 189

11.3 More Coherent Arguments 190

11.4 Coherency with Jaynes? 192

11.5 Coherency as $\boldsymbol{M}$, $\boldsymbol{N}$, and $\boldsymbol{n}$ Change 193

11.6 Causal Factors in Coin Tossing 198

12 Extending the Formal Rules for Prediction **203**

12.1 Conceptual Distinctions 203

12.2 Advancing the Bayesian paradigm 213

12.3 Bruno de Finetti's Representation Theorem 217

13 Predicting College Success **231**

13.1 Feller's Sample Space 231

13.2 Simple and Compound Events 236

13.3 Probability of Accidents During the Week 240

13.4 The Birthday Paradox 247

13.5 Bose–Einstein Statistics 251

13.6 Back to Predicting Success in College 264

14 Predicting College Success When Data Are Available **275**

14.1 The No Data Calculations 275

14.2 Data Become Available 279

14.3 Ratio of Posterior Probabilities for Models 291

14.4 Averaging Over All Models 292

15 What Does Uninformed Mean? 295

15.1 Extending the Kangaroo Scenario 295

15.2 Uninformed About What? 305

15.3 Pólya's Urn Scheme 307

16 Predicting the Behavior of Cellular Automata? 325

16.1 Computational Irreducibility 325

16.2 One Model is Deduced 327

16.3 Searching for New Physics 333

16.4 Probability Predictions 339

A Deconstructing *Mathematica* Code for Cellular Automata 343

A.1 The Motivation 343

A.2 The Goal 344

A.3 The Deconstruction (the easier part) 344

A.4 The Deconstruction (the harder part) 347

A.5 Adding My Own Obfuscations 351

B Brown's Function As A Cellular Automaton 355

B.1 Introduction 355

B.2 Rearranging the Functional Assignment Table 356

B.3 An Appropriate CA 357

B.4 Constructing Tables with *Mathematica* 360

C Discrete Probability Distributions 363

C.1 Introduction 363

C.2 Hypergeometric Distribution 364

C.3 Beta Binomial Distribution 365

C.4 Pólya–Eggenberger Distribution 370

C.5 Pólya Distribution 374

D Syntax of *Mathematica* Symbolic Expressions **377**

D.1 Introduction 377

D.2 Nested Syntax 377

D.3 Examples 378

D.4 Logic Functions 379

D.5 Diversity of Appearance for a Head 380

References **383**

List of Figures

3.1 *The Rule 81 elementary cellular automaton evolving over one time step* 31

3.2 *The Rule 59 elementary cellular automaton* 33

3.3 *The Rule 59 elementary cellular automaton showing the translation to a binary number* 34

3.4 *A three color cellular automaton evolving over 25 steps* 41

4.1 *A Venn diagram illustrating an ordering relationship* 54

4.2 *A four cell joint probability table with numerical assignments inspired by the forward implication logic function* 56

5.1 *A symbolic joint probability table using Boolean operations* 60

6.1 *Joint probability table with probability assignments dictated by the information in Rule 110* 91

6.2 *A joint probability table with probability assignments dictated by the information in Rule 170* 94

6.3 *All twenty possible samples of size 3 from a population of six individuals. The superscript number indicates a named individual* 120

6.4 *Using the* Mathematica *Boolean operators on three binary statements to help construct a joint probability table for a non–binary variable* . 122

6.5 *The final* 2×4 *joint probability table ensuing from Figure 6.4* 123

6.6 *A joint probability table constructed according to the ordering dictated by* **`Tuples[{T, F}, 3]`** 124

7.1 *The joint probability table for the second logic puzzle* 127

7.2 *A joint probability table involving three models for the Life on Mars scenario following from* $L \to W$ 132

8.1 *A joint probability table for the* `NAND` *and* `AND` *logic functions* 141

8.2 *A joint probability table based on Rule 150 and its complement to enforce inferences about the color of cells that have probabilities of 0 or 1* 145

8.3 *A joint probability table based on a rule number for a more complicated CA involving* two *nearest neighbors. A model based on this rule and its complement allows inferences about the color of cells that have probabilities of 0 or 1* 151

8.4 **`RulePlot[ ]`** *diagram for the complicated CA* 152

8.5 *The evolution of a more complicated deterministic CA for 20 steps* . 153

11.1 *Contingency table for 10,000 fictitious coin tosses* 199

13.1 *All 36 elementary points in the sample space for Feller's dice example* 235

15.1 *A* Mathematica *plot of a beta distribution with parameters* $\alpha = 7$ *and* $\beta = 3$ 313

16.1 *Examining the data from an unknown CA reveals that it must be running according to Rule 110. However, the prediction at the next time step is not correct* 329

16.2 *Examining the initial data about our World suggested that it might be running according to Rule 110. Further data demanded revision of this first tentative "Physics." As data about our World continue to accumulate, further revisions to the tentatively accepted "Physics" are required* 334

16.3 *An ontological system represented as a CA evolving according to some random rule* 338

A.1 *A CA reproducing the truth table for the* **`Nand[ ]`** *logic function* . . 346

B.1 *An expansion of Table B.1 showing the colors used in the CA* 358

C.1 *Plot of the beta binomial distribution showing the approach to both a binomial distribution with* $\delta(q - 1/2)$ *and a Normal distribution with* $\mu = 50$ 368

List of Tables

1.1 *Establishing the equivalency between the minterm canonical formula and a shorter formula* 4

2.1 *The first part of the functional assignment table for the logic puzzle* . 19

2.2 *The second part of the truth table for the logic puzzle* 20

3.1 *Mendelson's ordering for the truth table compared to the standardized* **`Tuples[ ]`** *ordering that* Mathematica *imposes* 35

3.2 *Truth table for some unknown Boolean function* 47

3.3 *Functional assignment table for three functions* 48

6.1 *Translating the truth table for Wolfram's simplest axiom into the correct elementary cellular automata rule number* 95

6.2 *The probabilities for the number of Ready individuals in a sample of size* $n = 3$ *taken from a population of* $N = 6$ 120

8.1 *Translating the placement of 0s in the eight cell joint probability table into Wolfram's rule number for an ECA* 146

9.1 *The data pare down the space of potential rules acting as models* . . 167

13.1 *Decomposition of the total sum of 81 possible elementary points for the case where* $M = 4$ *balls are distributed over* $n = 3$ *cells* 234

13.2 *Decomposition of the total sum of 7776 possible elementary points for the case where one die is rolled five times* 238

13.3 *Decomposition of the total sum of 1,716 possible contingency tables* . 243

13.4 *Decomposition of the total sum of 823,543 elementary points in the sample space* 245

13.5 *The relationship between the elementary points in the sample space for $r = 3$ and $n = 3$ and the definition of Bose–Einstein statistics* . . 253

13.6 *Accounting for all of the elementary points in the sample space for $M = 5$ kangaroos and $n = 8$ traits* 259

13.7 *Comparison of the probabilities for events under Maxwell–Boltzmann and Bose–Einstein statistics* 261

13.8 *An accounting of the elementary points in the sample space for $M = 6$ students and a state space of dimension $n = 8$* 265

13.9 *Translating a model into an ECA Rule Number* 268

15.1 *The probability of future frequency counts for two special eight cell contingency tables when the α_i parameters for the prior probability of a model are changed* 296

15.2 *A listing of 13,728 contingency tables from the total of nearly six billion that contain a frequency count of 73 kangaroos in one cell* . . 302

15.3 *The probability for five special contingency tables when all eight of the parameters for the prior probability approach 0* 307

15.4 *The probability for all four future frequency counts in the* Pólya Urn Scheme *sums to 1* 319

B.1 *Brown's original functional assignment table* 356

B.2 *My rearrangement of Brown's original table to correspond to the output from* **`Tuples[{T, a, a', F}, 2]`** *and the different notation in the carrier set* 357

Chapter 1

Boolean Algebra

1.1 Notation

These first five exercises start out by discussing and comparing Brown's [5] notation for Boolean Algebras and Boolean formulas within the context of the alternative notation employed throughout my Volume I. Brown's Example 3.7.1 on his page 41 is treated in depth.

Supplemental Exercise 1.1.1: Compare Brown's carrier set to mine.

Solution to Supplemental Exercise 1.1.1:

Brown defines one of his carrier sets as $\mathbf{B} = \{0, 1, a', a\}$, whereas I prefer to write it as $\mathbf{B} = \{a, a', F, T\}$.

Supplemental Exercise 1.1.2: Compare Brown's quintuple defining a Boolean Algebra to mine.

Solution to Supplemental Exercise 1.1.2:

Brown defines the quintuple $(\mathbf{B}, +, \cdot, 0, 1)$, whereas I write it as $(\mathbf{B}, \bullet, \circ, F, T)$.

Supplemental Exercise 1.1.3: Provide some examples of how Brown's function table is filled in.

Solution to Supplemental Exercise 1.1.3:

Brown presents this Boolean formula in his Table 3.3 on page 41 for the function,

$$f(x, y) = a'x + ay'$$

Brown uses the symbol "+" where I use the symbol "$\bullet$" for this binary operator in a Boolean Algebra. Brown's symbol "$\cdot$" is implicit in his formula and appears as my binary operator $\circ$. Thus, the function is re–written in my notation as,

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

Look at the seventh row in his Table 3.3. The first variable x is set to 1 and the second variable y is set to a'. The value of the function is $f(x = 1, y = a') = 1$. This result is confirmed in my notation by,

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

Turn now to the ninth row where the first variable x is set to a' and the second variable y is set to 0. The value of the function is $f(x = a', y = 0) = 1$. This result is confirmed in my notation by,

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

Turn now to the twelfth row where the first variable x is set to a' and the second variable y is set to a. The value of the function is $f(x = a', y = a) = a'$. This result is confirmed by,

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

Finally, look at the last row where the first variable x is set to a and the second variable y is also set to a. The value of the function is $f(x = a, y = a) = 0$. This

result is confirmed by,

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

Supplemental Exercise 1.1.4: How does Brown use Boole's Expansion Theorem for his example of a two variable Boolean function?

Solution to Supplemental Exercise 1.1.4:

Brown writes out the expansion of $f(x, y)$ in the reverse order from my presentation,

$$f(x, y) = x'y'\, f(0,0) + x'y\, f(0,1) + xy'\, f(1,0) + xy\, f(1,1)$$

as compared to,

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

He picks out the value for the four coefficient functions, which he calls *discriminants*, from his Table 3.3.

$$\begin{aligned} f(0,0) &= a \\ f(0,1) &= 0 \\ f(1,0) &= 1 \\ f(1,1) &= a' \end{aligned}$$

In my version this becomes,

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

With the coefficients determined, Brown's version of the function expansion is,

$$MCF(f) = ax'y' + xy' + a'xy$$

where Brown uses $MCF(f)$ to represent *Minterm canonical form* of the function f. My version is,

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

as discussed on pages 9–10 of Chapter One.

Brown does not discuss how to get from his minterm canonical form,

$$MCF(f) = ax'y' + xy' + a'xy$$

to the actual Boolean formula he uses in the example of Table 3.3,

$$f(x,y) = a'x + ay'$$

I began the process of demonstrating the logical equivalency between these two expressions in Volume I's Exercise 1.10.23.

Supplemental Exercise 1.1.5: Prove that the two formulas are equivalent.

Solution to Supplemental Exercise 1.1.5:

Brown points the way to a solution with a theorem he calls the *Löwenheim–Müller Verification Theorem*. The theorem says that we need only provide the T and F substitutions for the variables in order to verify a Boolean identity.

Table 1.1 below lists the four possible substitutions and the resulting function evaluations under the two different formulas. In each case they are the same, so the theorem allows us to claim that the formulas are identical. The final row is a check that equivalency still holds when another possible substitution is made from the carrier set that is not T or F.

Table 1.1: *Establishing the equivalency between the minterm canonical formula and a shorter formula.*

Case	(x,y)	$(a' \circ x) \bullet (a \circ y')$	$(a \circ x' \circ y') \bullet (x \circ y') \bullet (a' \circ x \circ y)$
1	TT	$(a' \circ T) \bullet (a \circ F) = a'$	$(a \circ F \circ F) \bullet (T \circ F) \bullet (a' \circ T \circ T) = a'$
2	TF	$(a' \circ T) \bullet (a \circ T) = T$	$(a \circ F \circ T) \bullet (T \circ T) \bullet (a' \circ T \circ F) = T$
3	FT	$(a' \circ F) \bullet (a \circ F) = F$	$(a \circ T \circ F) \bullet (F \circ F) \bullet (a' \circ F \circ T) = F$
4	FF	$(a' \circ F) \bullet (a \circ T) = a$	$(a \circ T \circ T) \bullet (F \circ T) \bullet (a' \circ F \circ F) = a$
5	aa'	$(a' \circ a) \bullet (a \circ a) = a$	$(a \circ (a' \circ a)) \bullet (a \circ a) \bullet (a' \circ (a \circ a')) = a$

1.2 Three Variable Boolean Functions

As in the previous exercises involving just two variable Boolean functions, let the carrier set consist of $\mathbf{B} = \{a, a', F, T\}$. Three variable Boolean functions are defined as $f : \mathbf{B} \times \mathbf{B} \times \mathbf{B} \to \mathbf{B} \equiv \mathbf{B^3} \to \mathbf{B}$. For example, $f(a', a, T) = F$.

Supplemental Exercise 1.2.1: Give a couple of examples from the *domain* of $\mathbf{B} \times \mathbf{B} \times \mathbf{B}$ expressed as $\mathbf{B^3}$.

Solution to Supplemental Exercise 1.2.1:

Any element from the set of ordered triples from $\mathbf{B^3}$ qualifies such as $\{a', a, T\}$ or $\{a, T, F\}$. A typical ordering might have the first element in the domain as $\{a, a, a\}$, with the last element as $\{T, T, T\}$. A complete listing is provided by,

```
Column[Table[Row[{i, Spacer[10], Tuples[{a, a', F, T}, 3]
                                                [[i]]}], {i, 1, 64}]]
```

Supplemental Exercise 1.2.2: How many elements are in the domain?

Solution to Supplemental Exercise 1.2.2:

There are four possibilities for each of the three slots standing for the x, y, and z variables. Thus, there $4 \times 4 \times 4 = 4^3 = 64$ possible elements in the domain. The value of 64 is returned by,

```
Length[Tuples[{a, a', F, T}, 3]]
```

Supplemental Exercise 1.2.3: Give a simple example of a formula for a three variable Boolean function.

Solution to Supplemental Exercise 1.2.3:

One function might be $f(x, y, z) = axyz$. This is an abbreviated form for,

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

Supplemental Exercise 1.2.4: What is the value of this function for the variable settings given above in Exercise 1.2.1?

Solution to Supplemental Exercise 1.2.4:

For the element in the domain given as a first example we have,

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

For the element in the domain given as a second example we have,

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

Supplemental Exercise 1.2.5: How many possible three variable functions exist for this carrier set?

Solution to Supplemental Exercise 1.2.5:

There are a total of $4^{4^3} = 4^{64} \approx 3.4 \times 10^{38}$ functions with three arguments for this Boolean Algebra.

Supplemental Exercise 1.2.6: Write down four functions that come to mind immediately.

Solution to Supplemental Exercise 1.2.6:

Two immediate functions from this huge total number of functions are the two constant functions $f(x, y, z) = F$ and $f(x, y, z) = T$. Having constructed these two functions, two more constant functions immediately come to mind, $f(x, y, z) = a$ and $f(x, y, z) = a'$.

Supplemental Exercise 1.2.7: Discuss a three variable Boolean function from the same perspective as taken in Supplemental Exercise 1.1.5.

Solution to Supplemental Exercise 1.2.7:

First of all, let's revert back to the simplest carrier set of $\mathbf{B} = \{F, T\}$. Consider now the three variable Boolean function,

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

What minterm canonical form, or for that matter, what fully expanded DNF, would reproduce such a function?

Things are simplified because the carrier set consists of only F and T, so the coefficients in the function expansion can only assume the values of T or F. Thus, either an orthonormal building block will appear if its coefficient is T or will be absent if its coefficient is F.

The generic template for the three variable function expansion is,

$$f(T, T, T)\, xyz \bullet \cdots \bullet f(F, F, F)\, x'y'z'$$

Five of the eight coefficients are T, while the three remaining coefficients are F,

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

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

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

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

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

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

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

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

Substituting into the full expansion template, retaining those five terms with a coefficient of T, while dropping those three terms with a coefficient of F, we have the minterm canonical form, or the full DNF, expressed as,

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

Expand each of the three terms in the original formula,

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

into three variables instead of just two to yield six terms,

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

Notice that the second and third terms are the same so we can drop one of them to see that the two formulas are equivalent,

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

Supplemental Exercise 1.2.8: Explain how the same function can be generated from a larger carrier set.

Solution to Supplemental Exercise 1.2.8:

Given the remarkable fact that we need only examine the functional values when the variable assignments are T or F, suppose that we restrict ourselves to just a two variable function $f(y,z)$ accompanied however by a larger carrier set consisting of $\mathbf{B} = \{F, x, x', T\}$. We now need only look at the four coefficients which might

take on the following legitimate values from the carrier set,

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

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

$$f(F,T) = x'$$

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

The minterm canonical function expansion then looks like,

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

Expand out the first term as was done in the previous exercise,

$$yz = xyz \bullet x'yz$$

and substitute,

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

to yield the same five terms in the minterm canonical expansion for,

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

The last two exercises were taken from Brown's treatment as presented on pp. 66–69. He wished to emphasize the notion of "Big Boolean Algebras" where the carrier set is not restricted to just F and T. For us, the significance of a "Big Boolean Algebra" is that it points the way to probability assignments for joint statements in the state space that are analogous to the coefficients (or *discriminants* as Brown calls them) in the function expansion.

1.3 Boole's Expansion Theorem for Three Variable Functions

These next exercises concentrate on the ramifications of Boole's Expansion Theorem for three variable functions. We gain some insight into how to think recursively.

Supplemental Exercise 1.3.1: What does Boole's Expansion Theorem prescribe for a three variable function?

Solution to Supplemental Exercise 1.3.1:

Let $f\colon \mathbf{B^3} \to \mathbf{B}$ be a Boolean function. Expand the function with respect to x,

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

Supplemental Exercise 1.3.2: Think recursively by again applying the theorem to $f(T, y, z)$, part of the first term on the right hand side of the expansion in the previous exercise.

Solution to Supplemental Exercise 1.3.2:

$f(T, y, z)$ is also a function of three variables, and it can be expanded in just the same way with respect to the y variable,

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

Now substitute this into the previous expansion,

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

At this point we have for the first term $[\ x \circ f(T, y, z)\ ]$,

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

Supplemental Exercise 1.3.3: Again, think recursively by applying the theorem to the $f(F, y, z)$, part of the second term on the right hand side of the initial expansion.

Solution to Supplemental Exercise 1.3.3:

$f(F, y, z)$ is also expanded with respect to the y variable,

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

Complete the second term $[\ x' \circ f(F, y, z)\ ]$ by,

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

At this point we have for the second term $[\ x' \circ f(F, y, z)\ ]$,

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

Supplemental Exercise 1.3.4: Substitute these two results as just found into the original expansion.

Solution to Supplemental Exercise 1.3.4:

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

Supplemental Exercise 1.3.5: Complete the decomposition of $f(x, y, z)$ by invoking the final recursive step.

Solution to Supplemental Exercise 1.3.5:

Each of the current four coefficient functions $f(\star, \star, z)$ will be expanded with respect to the z variable into two more functions through the same tactic as employed above. Thus, we arrive at the requisite eight terms for the full decomposition of the function using Boole's Expansion Theorem. The first term,

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

leads to,

$$f(T, T, T)\, xyz \bullet f(T, T, F)\, xyz'$$

The second term,

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

leads to,

$$f(T, F, T)\, xy'z \bullet f(T, F, F)\, xy'z'$$

The third term,

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

leads to,

$$f(F, T, T)\, x'\, yz \bullet f(F, T, F)\, x'yz'$$

and the fourth term,

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

leads to,

$$f(F, F, T)\, x'\, y'z \bullet f(F, F, F)\, x'y'z'$$

Putting everything back together, and reordering gives what we had hoped for,

$$\begin{aligned} f(x, y, z) \quad = \quad & f(T, T, T)\, xyz \bullet f(T, T, F)\, xyz' \bullet f(T, F, T)\, xy'z \bullet f(T, F, F)\, xy'z' \bullet \\ & f(F, T, T)\, x'yz \bullet f(F, T, F)\, x'yz' \bullet f(F, F, T)\, x'y'z \bullet f(F, F, F)\, x'y'z' \end{aligned}$$

Chapter 2

Logic Functions

2.1 Jaynes's Plausible Reasoning

I suppose that my own foray into Boolean Algebra and Classical Logic was partially driven by the emphasis that Jaynes placed on these topics very early on in his book. Jaynes had established a very compelling motivation that "plausible reasoning" was intended to generalize deduction. In order to understand what in fact was going to be generalized, Jaynes presented a very nice and concise introduction to Boolean Algebra and Logic in his Chapter 1.

It's sometimes difficult for an author to remember exactly to whom he should attribute credit for his own thoughts. In my case, both Jaynes and Wolfram played a significant role in how I chose to present my own version of logic functions in Chapter Two of Volume I.

So, a brief summary of our notational similarities and dissimilarities seems in order here. We both use A, B, C, ... for what Jaynes, following the traditional nomenclature, calls "propositions," and which I prefer to call "statements." At the outset though, given the abstract nature of Boolean Algebra, it seemed more appropriate to adhere to the x, y, z notation as introduced in my Chapter One. I switched over to the A, B, C, ... notation when transitioning from Boolean Algebra to Classical Logic in my Chapter Two.

Jaynes chose to represent the Boolean $\circ$ operation and the logic function $\wedge$ in an abbreviated form as AB, which I do as well. But I often times wish to reemphasize the functional nature of this expression and revert to the more correct $x \circ y$ or $A \wedge B$. For the disjunction, Jaynes writes $A + B$, while I prefer $x \bullet y$ or $A \vee B$. We both use the "overbar" to represent the denial of a proposition, or to indicate that a statement is FALSE as in $\overline{A}$. The prime symbol ($\prime$) was used in Boolean Algebra for the complement of an element in the carrier set $\mathbf{B}$ as in $(T)' = F$.

Jaynes rightfully points out that caution must be exercised with the use of the overbar symbol. For example, $\overline{AB}$ is different than $\overline{A}\,\overline{B}$. The first expression says that, "It is FALSE that both A and B are TRUE," while the second expression says that, "A is FALSE and B is FALSE." The first expression covers *three* possibilities, (1) A is TRUE and B is FALSE, (2) A is FALSE and B is TRUE, and (3) A is FALSE and B is FALSE, where we see that this third possibility is the one case covered by the second expression.

From the duality principle of Boolean Algebra, we will soon find out that this distinction translates over into a formal probability manipulation rules like,

$$P(\overline{AB}) = 1 - P(AB)$$

$$P(\overline{\overline{AB}}) = 1 - (1 - P(AB)) = P(AB)$$

$$P(A \vee B) = 1 - P(\overline{A} \wedge \overline{B})$$

Jaynes then presents a brief summary of the five main axioms from Boolean Algebra labeling them the same way as I have done (This seems to be the generally accepted nomenclature). I prefer the luxury of a more copious set of axioms for Boolean Algebra so my eventual list is longer. The five he listed are **Idempotence**, **Commutativity**, **Associativity**, **Distributivity**, and **Duality**. The **Duality Principle** and **De Morgan's axiom** are closely related.

Supplemental Exercise 2.1.1: What does Jaynes say about *Implication*?

Solution to Supplemental Exercise 2.1.1:

The very next topic Jaynes addresses is *Implication*. This is indeed a very confusing issue to anyone first exposed to Classical Logic. It could be that a lot of students drop the subject right there and then, when something so seemingly straightforward as the meaning for the word *implication* becomes fraught with technicalities. I would guess that those who do decide to stick it out must resolve the issue in ways that appeal to personal idiosyncracies. Jaynes has his way of talking around the issue and I have mine.

We will get to Jaynes in a second, but my own personal resolution for *implication* is this: Any attempts at verbal gyrations to somehow quasi–convince myself of the meaning of the word are doomed. In the end, I decided to just accept it for what it is. It is one defined logic function from the total possible number of 16 logic functions. Every one of these is clearly defined, and the `IMPLIES` operator is not any more mysterious than the other 15 logic functions when this viewpoint is taken.

Therefore, since only one variable assignment of $A = T$ and $B = F$ results in the functional assignment of F, the other three possible variable assignments result in T. If $A = F$ and $B = T$, I accept that the functional assignment is T. I don't try to hurt my brain cells by wondering how the word *implies* could possibly justify the situation where the premise is FALSE and the conclusion is TRUE.

Jaynes inserts far more content into his personal resolution, together with far more ramifications down the road for the development of probability theory as a generalization of logic, than my bald acceptance of an abstract definition. He says [11, pg. 11],

> The proposition
>
> $$A \Rightarrow B$$
>
> to be read as 'A implies B', does not assert that either A or B is true; it means only that $A\overline{B}$ is false, or, what is the same thing, $(\overline{A} + B)$ is true. This can be written as the logical equation $A = AB$. That is, given [the implication], if A is true then B must be true, or if B is false, then A must be false. This is just what is stated as the strong syllogism ...
>
> On the other hand, if A is false, [the implication] says nothing about B: and if B is true, [the implication] says nothing about A. But these are just the cases in which our weak syllogisms ... do say something. In one respect, then, the term 'weak syllogism' is misleading. The theory of plausible reasoning based on weak syllogisms is not a 'weakened' form of logic; it is an *extension* of logic with new content not present at all in conventional deductive logic.

Supplemental Exercise 2.1.2: Verify Jaynes's assertion in the above quote about *Implication* that $(\overline{A} + B)$ is true through the Duality Principle.

Solution to Supplemental Exercise 2.1.2:

Volume I's Table 2.6, listing all 16 logic functions, was generated in a completely systematic way by examining all possible ways of making a T or F the functional assignment. The listing started with all four Fs and ended with all four Ts.

Eventually, the `IMPLIES` logic function $f_{13}(A, B)$ was reached which had the functional assignment T matched up with the variable assignments TT, F with TF, T with FT, and T with FF. This meant that the disjunctive normal form (DNF) for implication was,

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

So we can see with $f(T, F) = F$ where "$A\overline{B}$ is false" comes from. The duality principle would allow us to then assert that,

$$A \wedge \overline{B} = F \longrightarrow \overline{A} \vee B = T$$

and we see in turn where "$(\overline{A} + B)$ is true" comes from.

Supplemental Exercise 2.1.3: Verify the assertion that $\overline{A} \vee B$ follows from the implication logic function through another Boolean approach.

Solution to Supplemental Exercise 2.1.3:

First,

$$\overline{A} = \overline{A}B \vee \overline{A}\,\overline{B}$$

$$B = AB \vee \overline{A}B$$

These steps are justified by,

$$B \vee \overline{B} = T$$

$$T \wedge \overline{A} = \overline{A}$$

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

$$\overline{A} = \overline{A}B \vee \overline{A}\,\overline{B}$$

and similarly,

$$A \vee \overline{A} = T$$

$$T \wedge B = B$$

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

$$B = AB \vee \overline{A}B$$

which then allows us to proceed to,

$$\begin{aligned}\overline{A} \vee B &= \overline{A}B \vee \overline{A}\,\overline{B} \vee AB \vee \overline{A}B \\ &= AB \vee \overline{A}B \vee \overline{A}B \vee \overline{A}\,\overline{B} \\ &= AB \vee \overline{A}B \vee \overline{A}\,\overline{B}\end{aligned}$$

The right hand side is, of course, the DNF for the `IMPLIES` logic function.

Supplemental Exercise 2.1.4: What further interesting comments does Jaynes make with regard to *Implication*?

Solution to Supplemental Exercise 2.1.4:

Jaynes advises us to be careful concerning a "tricky point": [11, pg. 12]

> *A tricky point*
>
> Note carefully that in ordinary language one would take 'A implies B' to mean that B is logically deducible from A. But, in formal logic, 'A implies B' means only that the propositions A and AB have the same truth value. In general, whether B is logically deducible from A does not depend only on the propositions A and B; it depends on the totality of propositions ... that we accept as true and which are therefore available to use in the deduction.
>
> ... Obviously, merely knowing that propositions A and B are both true does not provide enough information to decide whether either is logically deducible from the other, plus some other unspecified 'toolbox' of other propositions. ... The great difference in the meaning of the word 'implies' in ordinary language and formal logic is a tricky point that can lead to serious error if not properly understood; it appears to us that 'implication' is an unfortunate choice of word, and that this is not sufficiently emphasized in conventional expositions of logic.

Supplemental Exercise 2.1.5: Rely on *Mathematica* to verify Jaynes's discussion of the `NAND` logic function.

Solution to Supplemental Exercise 2.1.5:

On page 16, Jaynes contends that the single logic function `NAND` by itself can replace the three logic functions, `AND`, `OR`, and `NOT` which we used as an adequate set of operations in the disjunctive normal form. He shows three equations involving just `NAND` that accomplish this replacement:

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

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

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

Use the **`TautologyQ[ ]`** built–in function to assess whether logical equivalency exists between the two sides of the above three equations.

```
TautologyQ[Equivalent[Nand[a, a], Not[a]]]

TautologyQ[Equivalent[Nand[Nand[a, b], Nand[a, b]], And[a, b]]]

TautologyQ[Equivalent[Nand[Nand[a, a], Nand[b, b]], Or[a, b]]]
```

Each evaluation returns **`True`** confirming Jaynes's equations.

While we are at it, we might as well ask *Mathematica* for `NAND`'s truth table,

```
BooleanTable[Nand[a, b]]
```

returning the list,

```
{False, True, True, True}
```

to confirm Jaynes's writing down $A \uparrow B \equiv \overline{AB} = \overline{A} + \overline{B}$. The one term in the DNF with a coefficient of F is $A \wedge B$. Through the **Duality Principle**, $\overline{A} \vee \overline{B} = T$. As expected, *Mathematica* evaluates,

```
BooleanConvert[Nand[a, b]] // TraditionalForm
```

as $\neg a \vee \neg b$, or in **FullForm** as **Or[Not[a], Not[b]]**.

2.2 Another Logic Puzzle

In Volume I's section 2.9, I presented my own version of the solution to a logic puzzle that had appeared in one of my references to Boolean Algebra [15]. Here is the solution to a second similar logic puzzle appearing as Mendelson's Solved Problem 1.17(b) on his page 25.

> If the budget is not cut, then a necessary and sufficient condition for prices to remain stable is that taxes will be raised. Taxes will be raised only if the budget is not cut. If prices remain stable, then taxes will not be raised. Hence taxes will not be raised.

Unfortunately, while Mendelson's solution to the first puzzle was correct, his solution to this second puzzle was incorrect. He claimed that the argument was a valid one, while, in fact, it is not as I intend to show in subsequent exercises.

It is curious fact that in the nearly fifty years since the incorrect solution to the puzzle appeared in print, the mistake has not been corrected. I wonder how many teachers and students have just accepted the author's verdict without doing any verifications. Of course, I have to admit that I also would have merrily gone on my way blindly accepting the mistake if I had not had access to *Mathematica*. It was only because *Mathematica* could check all the possibilities instantaneously that I even ventured to undertake a verification and discovered where the mistake occurred.

I have chosen to present the quick solution first. Only then do I drill down and expand all of the compressed components for a more thorough analysis. This effort provides several supplemental exercises on the topic of Classical Logic first introduced in Chapter Two.

Supplemental Exercise 2.2.1: Solve this puzzle using *Mathematica*.

Solution to Supplemental Exercise 2.2.1:

The premise and the conclusion are composed from these three statements:

Statement #1. $B \equiv$ "The budget will be cut."

Statement #2. $P \equiv$ "Prices will remain stable."

Statement #3. $R \equiv$ "Taxes will be raised."

The symbolic expression for the premise is then,

$$(\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \wedge (P \rightarrow \overline{R})$$

The symbolic expression for the conclusion is simply $\overline{R}$.

It is instructive to see how Mendelson converted everyday language into logical expressions. This initial conversion process might be the major hang up for people to get started on these kinds of puzzles. "If ... then" language is converted into an implication for the first three sentences. These separate sentences are joined by an `AND` operator in the symbolic expression ($\wedge$) to establish the premise. Mendelson translates the language "necessary and sufficient condition" in the first sentence into a `BICONDITIONAL` operator ($\leftrightarrow$). Alternative language for this operator, which we also called `EQUAL`, is "if and only if." The word "Hence" is the signal that the conclusion follows from the stated premises.

Let's proceed directly to the *Mathematica* implementation to see the solution before we make any further comments. We have to establish that a tautology exists when the premise and the conclusion are joined by the `EQUAL` operator. Within *Mathematica*, the Boolean function **`Equivalent[ ]`** or **`Xnor[ ]`** is how this goal is accomplished.

The template for testing whether a tautology exists then looks like this,

```
TautologyQ[Equivalent[le1, le2]]
```

where **`le1`** stands for the first logic expression representing the premise, and **`le2`** is the second logic expression representing the conclusion.

Begin the process by translating each sentence in the premise into what will eventually be combined as **`le1`**, the first logic expression.

$\overline{B} \rightarrow (P \leftrightarrow R) \equiv$ **`Implies[Not[b], Xnor[p, r]]`**
$R \rightarrow \overline{B} \equiv$ **`Implies[r, Not[b]]`**
$P \rightarrow \overline{R} \equiv$ **`Implies[p, Not[r]]`**

Wrap **`And[ ]`** around these constituent expressions to form **`le1`**,

```
le1 = And[Implies[Not[b], Xnor[p, r]], Implies[r, Not[b]],
          Implies[p, Not[r]]]
```

The conclusion is simply **`le2 = Not[r]`**.

Evaluate the template since both the premises and the conclusion have been implemented,

```
TautologyQ[Equivalent[le1, le2]]
```

Now, at this point I was fully expecting the evaluation to return **True** confirming Mendelson's stated solution that the argument leading from the premises to the conclusion was valid. The conclusion that taxes would not be raised was supposedly logically equivalent to the stated assumptions.

However, the tautology evaluation returned **False**. There commenced a long deconstruction on my part to find out where the error had occurred. And, rightly so, I naturally assumed that I had made a mistake somewhere along the way. But the deconstruction forced me to revisit many of the introductory notions of Chapter Two as detailed in the upcoming exercises.

Supplemental Exercise 2.2.2: Continue to rely on *Mathematica* to do some initial detective work.

Solution to Supplemental Exercise 2.2.2:

I began to scour the *Mathematica* documentation on Boolean functions for some assistance. I found a predicate function **SatisfiableQ[** *bf* **]** that asks if there is any combination of assignments of TRUE or FALSE that would make the tautology come to pass.

But first I created a new function,

```
le3[b_, p_, r_] := Equivalent[And[Implies[Not[b],
                      Xnor[p, r]], Implies[r, Not[b]],
                      Implies[p, Not[r]]], Not[r]]
```

Now this function can serve as the argument *bf* to **SatisfiableQ[** *bf* **]**, with,

```
SatisfiableQ[le3[b, p, r]]
```

returning **True**. So, there was at least *one* set of assignments that satisfied the argument. But, critically, the definition of logical equivalency and the existence of a tautology demands that *all* possible assignments yield **True**.

I needed to find a more informative built–in function that would help me to actually identify where the assignments yielded **True**, and since **TautologyQ[ ]** had returned **False**, there must be at least one such assignment that yielded a **False**. If I could find this assignment, then it would pinpoint where the error had occurred.

Fortunately for my cause, there exists **SatisfiabilityInstances[** *bf* **]** that attempts to find a choice of variable assignments that make the Boolean function **le3[b, p, r]** return **True**. Even better, there is an additional argument to **SatisfiabilityInstances[** *bf* **]** that finds *all* instances where the Boolean expression is **True**. This is exactly what was required.

```
SatisfiabilityInstances[le3[b, p, r], {b, p, r}, All]
```

identified *seven* instances where the function returned **True**. So, the argument was **almost** valid. Seven out of eight assignments yielded **True**. But the fly in the ointment was that one assignment yielded **False**.

Scanning the list of the seven variable assignments to **b**, **p**, and **r** that did yield a **True**, as shown below,

```
{{True, True, True}, {True, True, False}, {True, False, True},
{True, False, False}, {False, True, True}, {False, False, True},
{False, False, False}}
```

points the guilty finger at the assignment **{False, True, False}** missing from the above list.

Therefore, when the assignment $B = F$, $P = T$, and $R = F$ is made to the premise, the outcome does not match the outcome to the conclusion. There is no logical equivalency; there is no tautology; the argument is not a valid argument; Mendelson's stated solution that it is, in fact, a valid argument is mistaken.

Supplemental Exercise 2.2.3: Drill down to the level of all functional assignments to verify the above result.

Solution to Supplemental Exercise 2.2.3:

Table 2.1 at the bottom of the page shows the first part of the functional assignment table. The eight possibilities for the variable assignment are indexed by the first column. These are systematically assigned according to the ordering dictated by the *Mathematica* **Tuples[ ]** function. This is elaborated on in the next exercise.

Table 2.1: *The first part of the functional assignment table for the logic puzzle. The assignment in column 6 is constructed from columns 3 and 4, and so on.*

Case	B	$\overline{B}$	P	R	$\overline{R}$	$P \leftrightarrow R$	$\overline{B} \rightarrow (P \leftrightarrow R)$	$R \rightarrow \overline{B}$	$P \rightarrow \overline{R}$
	1	2	3	4	5	6 (3, 4)	7 (2, 6)	8 (4, 2)	9 (3, 5)
1	T	F	T	T	F	T	T	F	F
2	T	F	T	F	T	F	T	T	T
3	T	F	F	T	F	F	T	F	T
4	T	F	F	F	T	T	T	T	T
5	F	T	T	T	F	T	T	T	F
6	F	T	T	F	T	F	F	T	T
7	F	T	F	T	F	F	F	T	T
8	F	T	F	F	T	T	T	T	T

The assignments are made according to the logic expression appearing in the column heading. For example, under the seventh column heading of $\overline{B} \rightarrow (P \leftrightarrow R)$, the functional assignment of F is made for the sixth row where $B = F$, $P = T$ and $R = F$. The `EQUAL` operator $f_8(T, F)$ returns F. $P \leftrightarrow R$ is F. The `IMPLIES` operator with arguments of T and the just computed value of F from f_8 returns F, that is, $f_{13}(T, F) = F$.

Table 2.2 below shows the second part of the assignment table. The second column is the conjunction of the three premises. The third column is the conclusion. It is easy to see how the functional assignment comes about for the `EQUAL` operator in the last column by comparing the functional assignments in the second and third columns. All the cases result in T except for Case 6.

Table 2.2: *The second part of the truth table for the logic puzzle.*

Case	$(\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \wedge (P \rightarrow \overline{R})$	$\overline{R}$	$\leftrightarrow$
1	F	F	T
2	T	T	T
3	F	F	T
4	T	T	T
5	F	F	T
$\boxed{6}$	$\boxed{F}$	$\boxed{T}$	$\boxed{F}$
7	F	F	T
8	T	T	T

The details in these tables confirm that Case 6 where $B = F$, $P = T$, and $R = F$ is where the tautology fails. This is exactly the case that was highlighted by **`SatisfiabilityInstances[ ]`** as the suspect case for the failure of logical equivalency.

Supplemental Exercise 2.2.4: Use the functional notation for Boolean operators as introduced in Chapter Two on the previous exercise.

Solution to Supplemental Exercise 2.2.4:

For a numerical example, choose the failure mode for the tautology as discovered in the previous exercise. The assignments to the Boolean variables are $B = F$, $P = T$, and $R = F$. The Boolean operators we require are `AND`, `EQUAL`, `NOT`, and `IMPLIES`. The functional notation is,

$$\texttt{AND} \equiv f_5(A, B)$$

$$\texttt{EQUAL} \equiv f_8(A, B)$$

$$\texttt{NOT A} \equiv f_6(A, B)$$

$$\texttt{IMPLIES} \equiv f_{13}(A, B)$$

If you intend to follow along with me, prepare yourself for some tedium ahead. Working from the innermost expressions outwards, we have,

$$P \leftrightarrow R \equiv f_8(T, F) = F$$

$$\overline{B} \rightarrow (P \leftrightarrow R) \equiv f_{13}(f_6(F, T), F) = f_{13}(T, F) = F$$

$$R \rightarrow \overline{B} \equiv f_{13}(F, f_6(T, T)) = f_{13}(F, F) = T$$

$$(\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \equiv f_5(F, T) = F$$

$$P \rightarrow \overline{R} \equiv f_{13}(T, f_6(T, T)) = f_{13}(T, F) = F$$

$$(\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \wedge (P \rightarrow \overline{R}) \equiv f_5(F, F) = F$$

$$(\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \wedge (P \rightarrow \overline{R}) \leftrightarrow \overline{R} \equiv f_8(F, f_6(F, T)) = f_8(F, T) = F$$

In words, the last line of the theorem says Premise `EQUAL` Conclusion is FALSE.

Supplemental Exercise 2.2.5: Employ *Mathematica* to double–check the entries in the assignments tables.

Solution to Supplemental Exercise 2.2.5:

Of course, if you find checking the assignments by hand too tedious, you can always have *Mathematica* do the job for you. Personally, I find hand checking kind of fun, a pleasant little mental exercise that forces one to concentrate. But it is imperative to have *Mathematica* double–check any work done by hand.

One may drill down to whatever level of the expressions in the premises or conclusions is desired, but for our purposes here let's check the two highest level expressions with **`BooleanTable[ ]`**.

First up is the conjunction of the premises, the heading of the second column in Table 2.2,

```
BooleanTable[And[Implies[Not[b], Xnor[p, r]],
                 Implies[r, Not[b]], Implies[p, Not[r]]]]
```

which returns the list,

```
{False, True, False, True, False, False, False, True}
```

matching the hand computed entries in the second column of Table 2.2.

Next, check the entries under the last column containing the logical equivalency between the premises and the conclusion.

```
BooleanTable[Equivalent[And[Implies[Not[b], Xnor[p, r]],
            Implies[r, Not[b]], Implies[p, Not[r]]], Not[r]]]
```

which returns the list,

```
{True, True, True, True, True, False, True, True}
```

matching the hand computed entries in the final column and, moreover, verifying that a tautology does not exist since the list does not exclusively contain **`True`**.

Notice that the Boolean operator **`Xnor[ ]`** is used for the first statement in the premise instead of **`Equivalent[ ]`** to mark a distinguishable break for the testing of equivalency between the premises and the conclusion.

Supplemental Exercise 2.2.6: Now that we know where the argument is not valid, try to pinpoint the source of the confusion.

Solution to Supplemental Exercise 2.2.6:

Recall that the failure mode occurs at the variable settings of $B = F$, $P = T$, and $R = F$. The "if and only if" clause $P \leftrightarrow R$ "Prices remain stable if and only if taxes are raised." is FALSE at the setting of $P = T$ and $R = F$, or $f_8(T, F) = F$. The implication "If ... then" $\overline{B} \rightarrow (P \leftrightarrow R)$ "If the budget is not cut, then prices remain stable if and only if taxes are raised." also is FALSE. Remember that for `IMPLIES` as a Boolean operator, $f_{13}(T, F)$ returns F. The remaining two sentences in the premises are irrelevant because the first sentence we have just examined is FALSE, and the entire premise is wrapped up in the `AND` operator, so the premise must be FALSE.

If a tautology is to exist, the conclusion must also be FALSE. The `EQUAL` operator will assign a T if either of the arguments match. The conclusion is $\overline{R} = T$, "taxes will be raised." creating a mismatch for the `EQUAL` operator, $f_8(F, T) = F$, and destroying any chance for the existence of a valid argument.

I think at this point we can identify the source of the confusion. We must clearly distinguish between a variable assignment of F and any binary operation or application of one of the sixteen logic functions. Strict adherence to the operator table of Table 2.5, page 39, Volume I, shows that the Boolean operator NOT A with

arguments F and T must return $f_6(F,T) = T$. Thus, as all possible assignments were rotated through in order to check for a tautology, and specifically when $R = F$ as a variable assignment was reached, the binary operator of negation turned this into a T.

A Boolean formula is given in terms of letters A, B, C, $\cdots$, and binary operators as in,

$$(\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \wedge (P \rightarrow \overline{R})$$

but, to be very clear, it should be specified what value from the carrier function **B** each letter is assuming.

Supplemental Exercise 2.2.7: Discuss some of the confusion surrounding the notation for the Classical Logic function of negation.

Solution to Supplemental Exercise 2.2.7:

Both my treatment in Chapter Two and the *Mathematica* implementation agree on the interpretation of a negation. It was simply in the interests of a concise notation that the "overbar" was placed on a statement as in $\overline{A}$ to indicate that statement A was FALSE. In this case, it would have been better to stick to the same notation used for all of the other functions in order to alleviate the confusion surrounding a negation evident in this logic puzzle.

All sixteen Classical Logic functions are defined over two statements, where each statement can only take on the values of T or F as was shown in Volume I's Table 2.5. We used functions like A `AND` $B \equiv A \wedge B$, or A `IMPLIES` $B \equiv A \rightarrow B$ with no confusion over the functional assignments to all four possible statement assignments.

Alternatively, the notation $f_5(T,F) = F$ or $f_{13}(F,F) = T$ did not allow for any ambiguity. The negation of statement A is also clearly defined under either notation A `NOT A` $B \equiv A \vdash B$, or $f_6(T,T) = F$ and $f_6(F,T) = T$. The setting of the second argument is immaterial since $f_6(T,F)$ is also F and $f_6(F,F)$ is also T.

We are also quite familiar with the *Mathematica* notation for some of the more common Classical Logic functions such as $A \rightarrow B \equiv$ **`Implies[p, q]`** or $A \leftrightarrow B \equiv$ **`Xnor[p, q]`**. In Figure A.1 of Volume I's Appendix A, I presented a very detailed mapping of Wolfram's definition and numbering of the sixteen Classical Logic functions to the ones given in Chapter Two.

From this diagram, we can see that the `NOT A` operator was represented by the binary number 0011, or the third Boolean function of two variables. So, what we wrote as $A \vdash B$ to perform the negation of A is in *Mathematica*,

```
f = BooleanFunction[3, 2]
BooleanConvert[f[p, q]] // FullForm
```

which returns **`Not[p]`**. Not unexpectedly, **`f[True, True]`** and **`f[True, False]`** both evaluate to **`False`**, while **`f[False, True]`** and **`f[False, False]`** both evaluate to **`True`**.

Supplemental Exercise 2.2.8: In preparation for a future exercise solving this puzzle using probability, find the full non–compressed disjunctive normal form of the premises. The DNF will then be used as the basis for the information in a model that assigns probabilities.

Solution to Supplemental Exercise 2.2.8:

We illustrate once again that probability theory when used to make inferences will be able to generalize any result from Classical Logic. Here this means that what deduction finds as a certainty, probability theory will replicate that result with a probability of 1.

The fully expanded, non–compressed version of the DNF, sometimes called the *minterm canonical form* or the *sum of products* (SOP) form, can be found by using **`BooleanTable[ ]`** on the premises, and then selecting only those terms that evaluate to T. This is the heuristic that we employed so often in Volume I.

The premises once again as expressions in Classical Logic look like this,

$$(\overline{B} \to (P \leftrightarrow R)) \wedge (R \to \overline{B}) \wedge (P \to \overline{R})$$

Translating over into *Mathematica*,

```
BooleanTable[And[Implies[Not[b], Xnor[p, r]],
                 Implies[r, Not[b]], Implies[p, Not[r]]]]
```

which returns the list,

```
{False, True, False, True, False, False, False, True}
```

This answer confirms the second column in Table 2.2.

Select the three **`True`** elements from the above list and match to the variable settings in order to construct the full DNF for the premises. We have for the first term, Case 2,

$$B = T, P = T, R = F, \text{ or } (B \wedge P \wedge \overline{R})$$

and for the second term, Case 4,

$$B = T, P = F, R = F, \text{ or } (B \wedge \overline{P} \wedge \overline{R})$$

and for the third term, Case 8,

$$B = F, P = F, R = F, \text{ or } (\overline{B} \wedge \overline{P} \wedge \overline{R})$$

The disjunction of these three terms becomes,

$$(B \wedge P \wedge \overline{R}) \vee (B \wedge \overline{P} \wedge \overline{R}) \vee (\overline{B} \wedge \overline{P} \wedge \overline{R})$$

or,

$$\text{DNF of premises } = BP\overline{R} \vee B\overline{P}\,\overline{R} \vee \overline{B}\,\overline{P}\,\overline{R}$$

When we ask *Mathematica* for the DNF of the premises through the auspices of **BooleanConvert[ ]**,

```
BooleanConvert[And[Implies[Not[b], Xnor[p, r]],
        Implies[r, Not[b]], Implies[p, Not[r]]]] // FullForm
```

we receive the compressed DNF,

```
Or[And[b, Not[r]], And[Not[p], Not[r]]]
```

which, despite its different appearance, is the same as the fully expanded DNF we found above.

How do we know that? Check to see whether a tautology exists between the two forms.

```
TautologyQ[Equivalent[
            Or[And[b, Not[r]], And[Not[p], Not[r]]],
            Or[And[b, p, Not[r]], And[b, Not[p], Not[r]],
               And[Not[b], Not[p], Not[r]]]]]
```

And guess what? The evaluation returns **True**.

This equality of forms can also be proved with somewhat more difficulty by resorting to the Boolean operations. Duplicate the second term, which doesn't change the Boolean expression, to set up for factoring,

$$\begin{aligned} BP\overline{R} \vee B\overline{P}\,\overline{R} \vee \overline{B}\,\overline{P}\,\overline{R} &= BP\overline{R} \vee B\overline{P}\,\overline{R} \vee B\overline{P}\,\overline{R} \vee \overline{B}\,\overline{P}\,\overline{R} \\ &= ((P \vee \overline{P}) \wedge B\overline{R}) \vee ((B \vee \overline{B}) \wedge \overline{P}\,\overline{R}) \\ &= B\overline{R} \vee \overline{P}\,\overline{R} \end{aligned}$$

2.3 Revisiting the First Logic Puzzle

Even though Mendelson's answer to his first logic puzzle was verified in section 2.9 of Volume I, the ensuing discussion of his solution was presented in a somewhat disorganized and haphazard fashion. At that early stage, the looseness was perhaps excusable. However, with the effort expended in solving his second logic puzzle, we can present the entire argument more concisely and more rigorously. In order to accomplish this goal, we will utilize the same *Mathematica* code developed in the last section.

Supplemental Exercise 2.3.1: Rework the first logic puzzle with the help of *Mathematica* following the same approach used above.

Solution to Supplemental Exercise 2.3.1:

First, develop the Boolean function expressing the potential tautology between the premise and the conclusion.

```
logicExpression[a_, b_, c_]:= Equivalent[
                                And[Or[Not[a], b, Not[c]],
                                Implies[Not[b], Or[a, c]]], b]
```

Secondly, ask *Mathematica* whether the expression is, in fact, a tautology with the predicate function,

```
TautologyQ[logicExpression[a, b, c]]
```

which returns **False** confirming our initial finding that the potential tautology in the puzzle failed.

Third, find out exactly where the tautology fails starting with,

```
SatisfiabilityInstances[logicExpression[a, b, c], {a, b, c}, All]
```

returning the list where six of the eight variable settings satisfied the tautology conditions with a functional assignment of **True**,

```
{{True, True, True}, {True, True, False}, {True, False, True},
{False, True, True}, {False, True, False}, {False, False, False}}
```

But it would be more informative if we could pick out those cases where the functional assignment of **False** was made,

```
Complement[Tuples[{True, False}, 3],
SatisfiabilityInstances[logicExpression[a, b, c], {a, b, c}, All]]
```

returning the list,

```
{{False, False, True}, {True, False, False}}
```

After these *Mathematica* evaluations, we are left with the satisfying outcome confirming our initial claim in section 2.9 that there are two settings of the variables

$$(1)\ A = F, B = F, C = T \text{ and } (2)\ A = T, B = F, C = F$$

where the tautology fails.

Supplemental Exercise 2.3.2: Rework Exercise 2.10.17 of Volume I with the help of *Mathematica* following the same approach used above.

Solution to Supplemental Exercise 2.3.2:

Volume I's Exercise 2.10.17 asked whether the following logic expression was a tautology,

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

First, define the logic expression in *Mathematica* with,

```
le[a_, b_, c_] := Equivalent[Xor[Implies[a, b], c]
                                        And[Or[a, b], Not[c]]]
```

If we then evaluate,

```
Reverse[Complement[Tuples[{True, False}, 3],
        SatisfiabilityInstances[le[a, b, c], {a, b, c}, All]]]
```

we get back the list of lists,

```
{{True, False, True}, {True, False, False}, {False, False, False}}
```

showing us the three variable settings where the tautology fails. The first variable setting in the above list was the one proven to result in the failure of the tautology in Exercise 2.10.17.

Chapter 3

Cellular Automata

3.1 ECA and the Second Logic Puzzle

We will tie in our initial explorations of logic functions and logic puzzles begun in Chapter Two with Wolfram's exposition of his 256 elementary cellular automata (ECA). These ECA are based on logic functions, or equivalently, Boolean functions. Three arguments from a carrier set lead to a functional assignment also from the carrier set,

$$f : \mathbf{B} \times \mathbf{B} \times \mathbf{B} \to \mathbf{B} \qquad \text{as in, for example, } f(T, F, T) = F$$

Here, the carrier set $\mathbf{B}$ consists of just the two special elements TRUE and FALSE.

These two elements are characterized as the colors black and white of any cell in the ECA. The ECA consists of successive applications of some Boolean function as given by some rule number where the three arguments are the colors of three cells at the previous times step and the functional assignment is the updated color at the current time step. We examined a detailed application of Rule Number 110 as an introductory example of an ECA in Volume I's Chapter Two.

We also examined a convenient heuristic showing how the DNF was formed from any given rule number. We will rely more and more on just the straightforward invocation of the appropriate *Mathematica* Boolean operations such as,

`BooleanConvert[ ]` and **`BooleanTable[ ]`**

when we need to carry out similar analyses of the DNF in these supplemental exercises.

Let's begin our review of the ECA with the solution to the second logic puzzle as discussed in the last Chapter.

Supplemental Exercise 3.1.1: Which of Wolfram's elementary cellular automata was involved in the solution to the second logic puzzle as it was discussed in the last Chapter?

Solution to Supplemental Exercise 3.1.1:

Rule 81 is the same as the set of premises in the second logic puzzle. How was the rule number discovered? A **`BooleanTable[ ]`** applied to the premises,

$$(\overline{B} \to (P \leftrightarrow R)) \wedge (R \to \overline{B}) \wedge (P \to \overline{R})$$

resulted in,

```
{False, True, False, True, False, False, False, True}
```

Translate these results into the DNF of the premises. Whenever a T appears, there appears a 1 at the same location in the binary expansion. Whenever a F appears, there appears a 0 at the same location in the binary expansion. The binary number is then,

$$\begin{aligned} FTFTFFFT &\to 01010001 \\ &= (0 \times 2^7) + (1 \times 2^6) + (0 \times 2^5) + (1 \times 2^4) + \\ &\quad (0 \times 2^3) + (0 \times 2^2) + (0 \times 2^1) + (1 \times 2^0) \\ &= 64 + 16 + 1 \\ &= 81 \end{aligned}$$

Supplemental Exercise 3.1.2: Draw the diagram showing the evolution of Rule 81 from some starting configuration.

Solution to Supplemental Exercise 3.1.2:

Use the standard way of displaying an elementary cellular automaton with,

```
ArrayPlot[CellularAutomaton[81,
          {1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0}, 1], Mesh → True]
```

to have *Mathematica* diagram Rule 81 as an elementary cellular automaton evolving over just one time step. The initial configuration of black and white cells at time step 0 was specified with the list **`{1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0}`**.

From this starting configuration, all of the eight functional relations between orthogonal building blocks and the coefficients can be checked. Reading from left to right, select out blocks of three cells in the top row and verify that the appropriate

color appears in the bottom row. The first three cells in the top row are all black and the second cell in the bottom row is white confirming that the first orthogonal building block of ABC has a coefficient of F.

Moving over one cell to the right, we have black, black, white in the top row with output black. The orthogonal building block of $AB\overline{C}$ must then have the coefficient of T for Rule 81 which it does. Consult Figure 3.1 below for a summary verification.

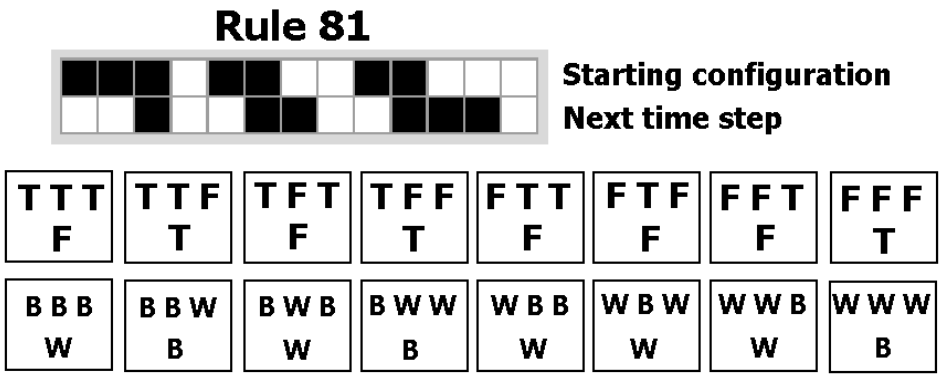

Figure 3.1: *The Rule 81 elementary cellular automaton evolving over one time step.*

Supplemental Exercise 3.1.3: What is Wolfram's logic expression for Rule 81?

Solution to Supplemental Exercise 3.1.3:

On page 884 of a *A New Kind of Science*, Wolfram lists the logic expressions for all 256 ECA. The expression for Rule 81 is the following,

$$(p \vee \neg q \vee r) \oplus r$$

Supplemental Exercise 3.1.4: Is this expression the same as the DNF expression?

Solution to Supplemental Exercise 3.1.4:

Yes it is. The fully expanded DNF expression is,

$$AB\overline{C} \vee A\overline{B}\,\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C} \equiv BP\overline{R} \vee B\overline{P}\,\overline{R} \vee \overline{B}\,\overline{P}\,\overline{R}$$

after employing the heuristic of finding the functional assignments of T in the above figure.

Form the *Mathematica* expression for this DNF as,

```
le1 = Or[And[p, q, Not[r], And[p, Not[q], Not[r]],
        And[Not[p], Not[q], Not[r]]]
```

and then form the *Mathematica* expression for Wolfram's expression as,

```
le2 = Xor[Or[p, Not[q], r], r]
```

Check on the tautology, or whether a logical equivalency exists between these two expressions, with,

```
TautologyQ[Equivalent[le1, le2]]
```

Since the evaluation returns **`True`**, this verifies that Wolfram's expression and my fully expanded DNF for Rule 81 are one and the same.

3.2 Rule 59 and the CNF

Up to this point, we have concentrated on one particular canonical form, namely, the disjunctive normal form (DNF). We have touted the DNF as being very helpful in transitioning from logic and deduction to inference and probability. There are many other canonical forms (you start to wonder what the word *canonical* is supposed to mean), and one of them is called the conjunctive normal form (CNF).

These two forms are highly symmetrical, with the source of the symmetry easy to understand because they are related by **De Morgan's axiom**. If there are more terms from Boole's Expansion Theorem with a coefficient of F as opposed to T, then it is more economical to develop the CNF.

This is a reworking within the context of Wolfram's ECA of a solved problem appearing in Mendelson (pg. 23) as 1.12(e). The problem as stated there is:

> Given a truth table for a truth function (not always taking the value T), construct a statement form in full cnf determining the given truth function.

Start by reinterpreting this problem as one where are trying to find an elementary cellular automaton that outputs black and white cells in a certain pattern at time step $N + 1$ given its three neighbors at the previous time step N. The pattern of black and white cells should match Mendelson's truth table. After discovering that this three variable Boolean function is none other than Rule 59, we can pose the question: What is the fully expanded conjunctive normal form for the elementary cellular automaton represented by Rule 59?

Supplemental Exercise 3.2.1: What does Rule 59 look like as one of the 256 elementary cellular automata?

Solution to Supplemental Exercise 3.2.1:

Display Rule 59 in the standard fashion with *Mathematica* through,

```
ArrayPlot[CellularAutomaton[59,
  {1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0}, 10], Mesh → True ]
```

as shown in Figure 3.2 below.

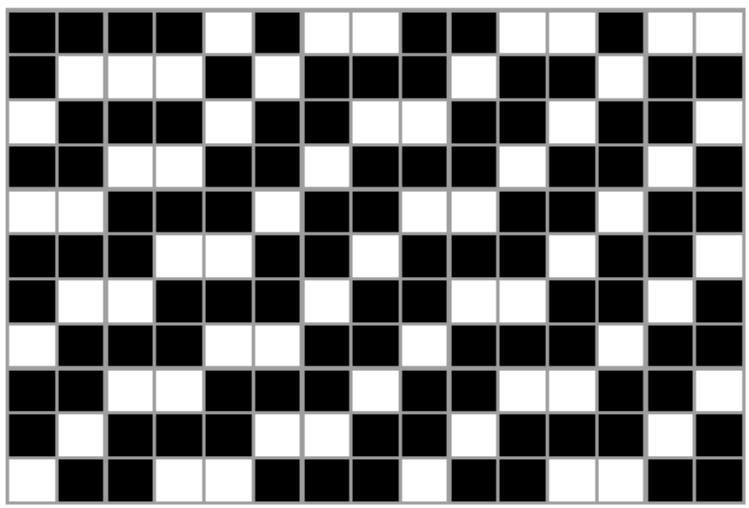

Figure 3.2: *The Rule 59 elementary cellular automaton.*

The list of 1s and 0s indicate the starting configuration of 15 white and black cells as run over 10 subsequent time steps. Examine this cellular automaton to observe that it produces a white cell if all three neighbor cells at the previous step were black. It produces a black cell if the neighboring cells were white, black, and black. And so on.

This pattern for Rule 59 is the same as the truth table for a particular logic function. Wolfram gives it in compressed form as,

$$(\neg p \wedge r) \vee (\neg q)$$

which for us would be,

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

Unfortunately, Wolfram's expression for Rule 59 is neither in disjunctive normal form nor in conjunctive normal form.

Supplemental Exercise 3.2.2: How is Rule 59 visually displayed for all eight possibilities of the neighboring cells?

Solution to Supplemental Exercise 3.2.2:

The top portion of Figure 3.3 shows the evolution of the cell color at each time step for Rule 59. It is entirely comparable to Figure 3.2 in Volume I showing the evolution of Rule 110. The bottom portion is the translation from the black and white cell color to the functional assignment for the three variable logic function $(\neg p \wedge r) \vee (\neg q)$. Wolfram's binary numbering system for the elementary cellular automata that results in labeling this ECA as Rule 59 follows.

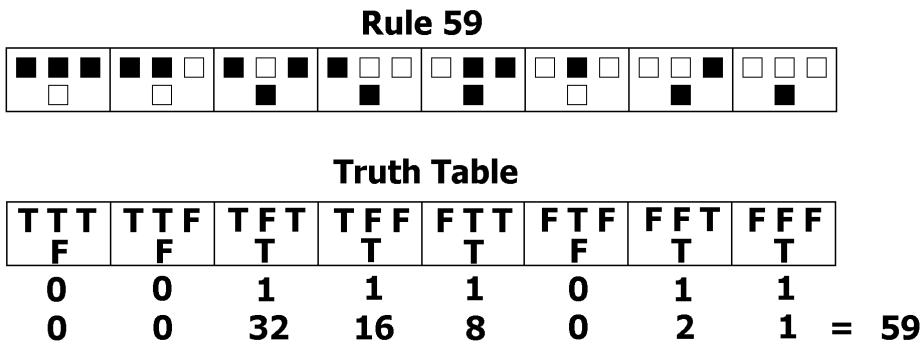

Figure 3.3: *The Rule 59 elementary cellular automaton showing the translation to a binary number.*

Mathematica will display the layout of any rule as shown in the top part of Figure 3.3 with **`RulePlot[CellularAutomaton[59]]`**.

Supplemental Exercise 3.2.3: Engage the standard set of *Mathematica* operations to check the results in the last exercise.

Solution to Supplemental Exercise 3.2.3:

Set the variable **`f`** to the rule 59 Boolean function,

```
f = BooleanFunction[59, 3]
```

Display the condensed DNF expression as either,

```
BooleanConvert[f[p, q, r]] // TraditionalForm
```

returning $(\neg p \wedge r) \vee \neg q$ or,

```
BooleanConvert[f[p, q, r]] // FullForm
```

returning **`Or[And[Not[p], r], Not[q]]`**.

Verify the bottom portion of Figure 3.3 with **`BooleanTable[f[p, q, r]]`** returning the list,

```
{False, False, True, True, True, False, True, True}
```

Supplemental Exercise 3.2.4: Complete the Mendelson example.

Solution to Supplemental Exercise 3.2.4:

Table 3.1 shows Mendelson's chosen order for the truth table so that it can be compared and the necessary correspondence made with the standardized order that *Mathematica* imposes through **`Tuples[ ]`** as illustrated above in Figure 3.3. The final column shows that the functional assignment $f_\star(A, B, C)$ for the *Mathematica* ordering is indeed the same.

Table 3.1: *Mendelson's ordering for the truth table compared to the standardized* **`Tuples[ ]`** *ordering that* Mathematica *imposes.*

Order	A	B	C	$f(A, B, C)$	**`Tuples[ ]`** order	$f_\star(A, B, C)$
1	T	T	T	F	1	**`False`**
2	F	T	T	T	5	**`True`**
3	T	F	T	T	3	**`True`**
4	F	F	T	T	7	**`True`**
5	T	T	F	F	2	**`False`**
6	F	T	F	F	6	**`False`**
7	T	F	F	T	4	**`True`**
8	F	F	F	T	8	**`True`**

Supplemental Exercise 3.2.5: What is the fully expanded conjunctive normal form for the elementary cellular automaton represented by Rule 59?

Solution to Supplemental Exercise 3.2.5:

This exercise is practice in expanding out Wolfram's Boolean expression for Rule 59. Be forewarned that this effort, like all of its brethren, is by its very nature tedious to the extreme. But to put you in a better frame of mind, the procedure is quite mechanical with none of the individual steps hard. As I have said before, I like this facet of Boolean Algebra because no operation is mysterious and no intuitive leaps are required at any point. It's what I wish all mathematics were like.

We were left with the expression $(\overline{A} \wedge C) \vee \overline{B}$ at the end of Exercise 3.2.1. Apply the **Distributivity axiom** in reverse to reach the intermediate step,

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

Now, "multiply" this expression once again relying upon the **Distributivity axiom** to arrive at,

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

Our objective is to get this expression into a fully expanded state where we can clearly see all three variables. Then, we can delete the repeated terms, and hopefully match up the remaining terms with the terms with coefficients of T from Boole's Expansion Theorem.

We start off by examining the above expression,

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

to see that expanding the first three terms will result in six terms, and expanding the final term will add four more terms for a total of ten terms.

These ten terms in the expansion are,

$$\overbrace{\overline{A}BC \vee \overline{A}\,\overline{B}C \vee \overline{A}\,\overline{B}C \vee \overline{A}\,\overline{B}\,\overline{C} \vee A\overline{B}C \vee \overline{A}\,\overline{B}C}^{6} \vee \overbrace{A\overline{B}C \vee \overline{A}\,\overline{B}C \vee A\overline{B}\,\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}}^{4}$$

Rearrange these terms so that the repeated terms appear next to each other,

$$\overline{A}BC \vee \overbrace{\overline{A}\,\overline{B}C \vee \overline{A}\,\overline{B}C \vee \overline{A}\,\overline{B}C \vee \overline{A}\,\overline{B}C}^{\checkmark} \vee \overbrace{\overline{A}\,\overline{B}\,\overline{C} \vee \overline{A}\,\overline{B}\,\overline{C}}^{\checkmark} \vee \overbrace{A\overline{B}C \vee A\overline{B}C}^{\checkmark} \vee A\overline{B}\,\overline{C}$$

After deleting the repeated terms, we are left with five terms in the fully expanded DNF, which is what we had hoped to find,

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

The order that we would ordinarily use for the DNF follows the *Mathematica* convention of,

$$\underbrace{A\overline{B}C}_{4} \vee \underbrace{A\overline{B}\,\overline{C}}_{5} \vee \underbrace{\overline{A}BC}_{1} \vee \underbrace{\overline{A}\,\overline{B}C}_{2} \vee \underbrace{\overline{A}\,\overline{B}\,\overline{C}}_{3}$$

the underbrace indicating the corresponding term from above.

Let's go ahead and verify these symbolic operations through the usual procedure of establishing a tautology between two expressions. The first expression is the fully expanded DNF consisting of five terms as just derived,

```
le3 = Or[And[p, Not[q], r], And[p, Not[q], Not[r]],
        And[Not[p], q, r], And[Not[p], Not[q], r],
        And[Not[p], Not[q], Not[r]]]
```

The second expression is the one returned by the **BooleanConvert[ ]** on Rule 59,

```
le4 = Or[And[Not[p], r], Not[q]]
```

The two expressions are confirmed as logically equivalent with,

```
TautologyQ[Equivalent[le3, le4]]
```

returning **True**.

Supplemental Exercise 3.2.6: Did we accomplish what we set out for ourselves in the last exercise?

Solution to Supplemental Exercise 3.2.6:

No, we did not. At the conclusion of the last exercise we had succeeded only in verifying the DNF for the Boolean function behind Rule 59. Rather than locating the five terms with a coefficient of T in the expansion, select instead the three terms with a coefficient of F. This is easy, and so we find the three terms,

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

The heuristic for constructing the CNF, following **De Morgan's axiom**, says that since we have found the terms with a coefficient of F as opposed to the terms with a coefficient of T, complement every statement variable, and switch $\wedge$ to $\vee$ and $\vee$ to $\wedge$,

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

This matches Mendelson's answer for the CNF in his chosen format of,

$$(\neg\mathbf{A} \vee \neg\mathbf{B} \vee \neg\mathbf{C})\,\&\,(\neg\mathbf{A} \vee \neg\mathbf{B} \vee \mathbf{C})\,\&\,(\mathbf{A} \vee \neg\mathbf{B} \vee \mathbf{C})$$

The role of the **And[ ]**s and **Or[ ]**s has been reversed, but the dual symmetry between the DNF and the CNF is evident. Let's make sure that we haven't messed up through,

```
FullForm[BooleanConvert[f[p, q, r], "CNF"]]
```

which returns **And[Or[Not[p], Not[q], Or[Not[q], r]]**. *Mathematica* gives us the option of specifying which kind of canonical form we want. The disjunctive normal form is the default. However, once again, *Mathematica* does not provide us with the fully expanded CNF, so we have to check for the existence of a tautology once again,

```
TautologyQ[Equivalent[le5, le6]]
```

where **le5** is the above expression returned by the **BooleanConvert[ ]** and **le6** is the fully expanded CNF as found above,

```
le6 = And[Or[Not[p], Not[q], Not[r]], Or[Not[p], Not[q], r],
                                      Or[p, Not[q], r]]
```

Supplemental Exercise 3.2.7: Confirm the diagram shown in Exercise 3.7.10 of Volume I with the standard *Mathematica* code for CA.

Solution to Supplemental Exercise 3.2.7:

We have been exposed to the standard *Mathematica* code to illustrate the evolution of elementary CA in previous supplemental exercises 3.1.2 and 3.2.1. With just a slight change to the syntax, we can easily confirm the five variable example of Volume I's Exercise 3.7.10.

The template for the first argument in **CellularAutomaton[ ]** is upgraded to a list {*rule number, number of colors, range*}. The third argument *range* is where the critical change takes place.

The *range* is defined as number of cells to the right or left of the cell to be updated, so instead of $r = 1$ as in the ECA, for this example with two neighbors on either side, it becomes $r = 2$. The *rule number* remains at 2,147,483,649, and the *number of colors* remains at 2.

Implementing these changes, evaluate,

```
ArrayPlot[CellularAutomaton[{2 147 483 649, 2, 2},
                 {1, 1, 1, 1, 1, 0, 0, 0, 0, 0}, 6], Mesh → True]
```

to confirm that the same visual output appears as in Volume I's Figure 3.5.

The initial configuration list matches the starting pattern of black and white cells shown in Figure 3.5 of Volume I of five black cells followed by five white cells. The CA is allowed to evolve for six time steps. Notice that creating the mesh is an *option* to **ArrayPlot[ ]**.

Supplemental Exercise 3.2.8: Engage in a pretty thorough discussion of five variable Boolean functions based on the cellular automaton of the last exercise.

Solution to Supplemental Exercise 3.2.8:

How many five variable logic functions $f(A, B, C, D, E)$ are there?

Using Wolfram's formula, we find that there are,

$$k^{k^n} = 2^{2^5} = 2^{32} = 4,294,967,296$$

possible logic functions for five variables, just as there were $2^{2^2} = 2^4 = 16$ logic functions $f(A, B)$ for two variables and $2^{2^3} = 2^8 = 256$ logic functions $f(A, B, C)$ for three variables.

Mathematica uses the slightly different notation of $k^{k^{(2r+1)}}$ for the number of rules. k has the same meaning denoting the number of colors for the cells of the CA, where r is the *range* for the number of neighbors on each side of the cell under consideration. Here then $r = 2$, where the two variables A and B are the left neighbors of variable C, while the two variables D and E are the right neighbors.

In Classical Logic, the carrier set **B** consists of only two ($k = 2$) elements, these being the so–called special elements of TRUE and FALSE. One function is the constant function $f(A, B, C, D, E) = T$ where every assignment is T. Another function is the constant function $f(A, B, C, D, E) = F$ where every assignment is F. Another function is a non–constant function where $f(A, B, C, D, E) = T$, $f(\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}) = T$, but every other assignment equals F.

What is the binary number attached to each of these logic functions?

Use **IntegerDigits[ ]** to find the coefficient for each of the $2^5 = 32$ building block functions 2^0 through 2^{31}. So, as expected,

```
IntegerDigits[4 294 967 295, 2, 32]
```

returns a list of 32 1s, and,

```
IntegerDigits[2 147 483 649, 2, 32]
```

also returns a list of 32 elements with the first and last a 1 and everything else a 0.

Pick any arbitrary rule number for a cellular automaton operating according to a five variable logic function. Say we chose rule number 3 916 273 040. We know that this function $f(A, B, C, D, E)$ will take on a functional assignment of T wherever **Integer Digits[ ]** reveals a 1, and, conversely, an assignment of F wherever **Integer Digits[ ]** reveals a 0 in the binary decomposition of the decimal expression for the rule number.

Proceed then to examine the output from,

```
IntegerDigits[3 916 273 040, 2, 32]
```

You may inspect the list of 32 1s and 0s to determine the functional assignment for any variable setting. For example, since the first element in the list as returned by **IntegerDigits[ ]** is a 1, then,

$$f_{3916273040}(A, B, C, D, E) = T$$

Since the last element in the list as returned by **IntegerDigits[ ]** is a 0, then,

$$f_{3916273040}(\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}) = F$$

Returning to the five variable logic function of Volume I and the last exercise, we know that the DNF for $f_{2147483649}(A, B, C, D, E)$ must be,

$$ABCDE \vee \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}\,\overline{E}$$

Have *Mathematica* verify this by first creating a Boolean function according to this rule number,

```
g = BooleanFunction[2 147 483 649, 5]
```

and then finding the DNF through,

```
BooleanConvert[g[a, b, c, d, e]] // FullForm
```

returning,

```
Or[And[a, b, c, d, e], And[Not[a], Not[b], Not[c], Not[d], Not[e]]]
```

The **ArrayPlot[ ]** written out in the previous exercise illustrates the cellular automaton operating according to this particular logic function of five variables.

3.3 Three Color CA

In order to orient ourselves for the upcoming exploration of three color CA, look at the slightly changed syntax as introduced in the last exercise for how *Mathematica* would produce a diagram of a three color CA following a very large rule number. The CA consists of yellow, blue, and red colored cells (rendered in shades of gray in the CA pictured at the top of the next page).

```
ArrayPlot[CellularAutomaton[{2 625 597 484 982, 3, 1},
    {0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 1, 0, 0, 1, 1, 0, 1, 2}, 25],
     ColorRules → {0 → Yellow, 1 → Blue, 2 → Red}, Mesh → True]
```

In the argument list {*rule number, number of colors, range*}, the number of colors has been changed from 2 to 3. The range remains at $r = 1$ since the color of the updated cell depends only on one left and right neighbor cell. The evolution of this three color CA is followed for 25 steps. The initial configuration list starts the CA off with five yellow cells, followed by a blue cell, followed by two more yellow cells, and so on, until at the final three cells there is a yellow, blue, and red cell.

See Figure 3.4 for what this three color CA looks like as it evolves over the first 25 steps. No simple repetitive structures appear early on, so it might be a candidate just like Rule 110 or Rule 30 as an interesting CA.

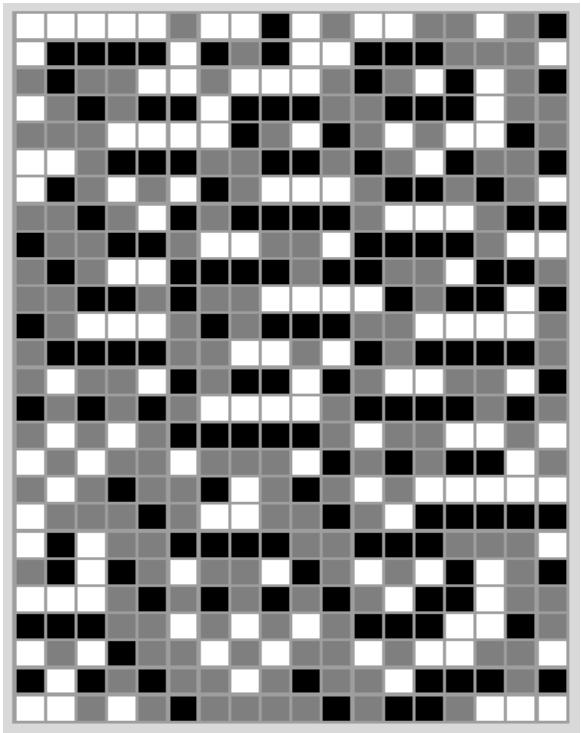

Figure 3.4: *A three color cellular automaton evolving over 25 steps. In the above rendering, Yellow appears as white, Blue as gray, and Red as black.*

The formula for the total number of rules in any cellular automaton consisting of k colors and a range r is,

$$\textbf{Total number of rules} = k^{k^{(2r+1)}}$$

The range is the number of cells being considered both to the right and left of the cell being updated, so for the case where we are still at three variables, $r = 1$. The formula then tells us that there is the enormous number of over seven trillion possible three color CA,

$$\textbf{Total number of rules} = 3^{3^{(2\times 1+1)}} = 3^{3^3} = 7,625,597,484,987$$

Just by increasing the number of colors from 2 to 3, we have increased the number of possible CA from $2^{2^3} = 256$ to the above number. We have another elementary example of a combinatorial explosion thwarting our mind's ability to visualize the situation. So it becomes very much harder to decode a rule number that could be as large as $7,625,597,484,986$ as opposed to 255.

It now becomes obvious why such a large number as 2 625 597 484 982 can appear as the rule number for a three color CA. After finding out the upper limit for the rule number, I just "randomly" punched enough numbers as a rule number so that I stayed below this limit, but still had a very large number.

Supplemental Exercise 3.3.1: How would you decode a rule number for a CA with three colors?

Solution to Supplemental Exercise 3.3.1:

How did we decode a rule number for a CA with two colors? For example, we looked at the binary decomposition of rule number 110 through,

```
IntegerDigits[110, 2, 8]
```

to retrieve the list **{0, 1, 1, 0, 1, 1, 1, 0}**. Then we could form,

$$(0\times 2^7)+(1\times 2^6)+(1\times 2^5)+(0\times 2^4)+(1\times 2^3)+(1\times 2^2)+(1\times 2^1)+(0\times 2^0)=110$$

showing the order of the eight color assignments, reading from left to right, as white, black, black, white, black, black, black, and white when matched with the colors of three cells at the previous time step: BBB, BBW, BWB, BWW, WBB, WBW, WWB, and finally WWW.

When we transition to base 3 number system, only the digits 0, 1, and 2 will appear. So, for example, the decimal number 51 is represented in base 3 as 1220,

$$(1\times 3^3)+(2\times 3^2)+(2\times 3^1)+(0\times 3^0)=51$$

and,

```
IntegerDigits[51, 3, 4]
```

returns the list **{1, 2, 2, 0}**.

For the 256 ECA, each of three cells could only take on one of two colors, black or white, leading to the specification for any rule number only when all eight possible color configurations had an assignment. Now, since each of three cells can take on any one of three colors, yellow, red, or blue, the specification for any three color rule number occurs only when all 27 possible color configurations receive an assignment.

Let's now decode our large three color rule number with,

```
IntegerDigits[2 625 597 484 982, 3, 27]
```

which returns a list of 27 0s, 1s, and 2s. The actual list begins and ends with,

```
{1, 0, 0, ···, 1, 2, 2}
```

Supplemental Exercise 3.3.2: How would you match up the decoding with what the CA produces as the color of the cell at the next time step?

Solution to Supplemental Exercise 3.3.2:

Start by examining the very last number in the list presented in the previous exercise by **`IntegerDigits[ ]`**,

```
{1, 0, 0, ···, 1, 2, 2}
```

The color represented by 2, in other words, the color red, must be the color of the updated cell when the relevant three cells at the previous time step are all yellow. Review the **`ColorRules`** option for **`ArrayPlot[ ]`** to verify the assignment of red to the digit "2".

The triple YYY (the three relevant cells at the previous time step are all yellow) is equivalent to an elementary CA consisting of only black and white cells where the last of the eight possibilities was always given as WWW. If a 1 were assigned to this particular color configuration by the specified rule number, then the updated cell was black, otherwise it could be white under a different rule number that assigned a 0.

Thus, the triple YYY, as the last of the 27 possibilities, leads to the updated cell being red. Verify this by examining the starting configuration of cell colors in Fig 3.4. The first three cells are all colored yellow, and the color of the second cell at the next step in the evolution is indeed red. In other words, the coefficient of the first term 3^0 from **`IntegerDigits[ ]`** is 2.

The next triple in order would be YYB, which according to the assignment by this rule number should update its cell also to red since the next number to the left is another 2. Find a sequence of yellow, yellow, blue cells in the starting configuration (fourth, fifth, and sixth cells) and sure enough the updated cell is red. The coefficient of the second term 3^1 from **`IntegerDigits[ ]`** is also 2.

Continuing on in the same vein, the next triple would be YYR. This should update to a blue cell as a 1 appears next in the list. Locate any yellow, yellow, red sequence in the evolution of the CA (seventh, eighth, and ninth in the starting configuration), to confirm that the updated cell at the next time step is blue. The coefficient of the third term 3^2 from **`IntegerDigits[ ]`** is 1.

Everything continues along in the same fashion. Jump ahead to the final triple of RRR from the total of 27. The coefficient of the final term 3^{26} found from **`IntegerDigits[ ]`** is also 1, therefore, the color assignment of the updated cell given that the above three cells were red, red, and red is blue. Verify this by locating any three red cells in sequence in Figure 3.4 to see that they produce a blue cell.

Supplemental Exercise 3.3.3: How would *Mathematica* encapsulate the discursive explanation of the last exercise?

Solution to Supplemental Exercise 3.3.3:

All of this is made very much easier by examining the output from,

```
RulePlot[CellularAutomaton[{2 625 597 484 982, 3, 1}],
        ColorRules → {0 → Yellow, 1 → Blue, 2 → Red}]
```

Here we see the 27 possibilities for all three colors of the relevant cells at the previous time step laid out in the order in which they would appear from an application of **`Tuples[ ]`**. They proceed in order from RRR leading to a blue cell all the way to the end where YYY leads to a red cell and everything in between.

Supplemental Exercise 3.3.4: Find the rule number for some three color three variable CA that is equivalent to a specified Boolean function with three arguments.

Solution to Supplemental Exercise 3.3.4:

To prime us for this task, review the analogy between the 256 ECA and Boolean functions. The expansion of a function consisting of three variables x, y, and z is,

$$f(x,y,z) = f(T,T,T)\,xyz \bullet \cdots \bullet f(F,F,F)\,x'y'z'$$

Let's review the Rule 110 ECA. The orthogonal building block for the first term, where we are now reading in the reverse direction from right to left, is WWW corresponding to $x'y'z'$. The color assignment is W, so the coefficient must be $f(F,F,F) = F$. The next term over is WWB corresponding to $x'y'z$. The color assignment is B, so the coefficient of this second term is $f(F,F,T) = T$.

Continuing on like this, we eventually reach the last term (the first term reading from left to right) xyz corresponding to BBB. The color assignment is W, so the coefficient of this last and eighth term is $f(T,T,T) = F$. The Boolean function analogous to the two color ECA Rule 110 is therefore,

$$f(x,y,z) = (F \circ xyz) \bullet \cdots \bullet (T \circ x'y'z) \bullet (F \circ x'y'z')$$

$$= xyz' \bullet xy'z \bullet xy'z' \bullet xy'z' \bullet x'y'z$$

Enlarge the carrier set to $\mathbf{B} = \{F, a, a', T\}$ in order to accommodate the new color red $a = a'$ in addition to black T and white F.

According to Brown [5], the truth table for any Boolean function of n variables consists of 2^n rows for all possible assignments of T and F as arguments. For this three variable example, we will therefore construct an $2^3 = 8$ row truth table.

The functional assignment $f(x, y, z)$ at each of these eight rows will determine the Boolean function.

The first row lists the three arguments for $f(x, y, z)$ as T, T, T, the second row has T, T, F, and so on to final eighth row with F, F, F. Select the functional assignments of $f(x = T, y = T, z = T) = T$ for the first row, $f(x = F, y = T, z = T) = a$ for the fifth row, with all remaining rows having a functional assignment of F. This was done to achieve the goal of a relatively simple Boolean function minterm canonical form expansion looking like this,

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

Now, let's translate all of this over to a three color CA with a range of $r = 1$. In other words, we will construct a CA equivalent to this Boolean function consisting of cell colors black, white, and red with the color of the updated cell depending on the previous cell and its two immediate neighbors.

We know the range of rule numbers for all possible CA of this type. We must figure out which rule number will duplicate the Boolean function. This will be a list of 27 digits all either 0, 1, or 2 with 0 indicating a white cell, 1 a red cell, and 2 a black cell. I first wrote down all 27 arrangements of the colors of the three cells at time step N and computed the resulting color at time step $N + 1$ from the Boolean function.

For example, the first arrangement of colors at time step N is BBB. This would correspond to the arguments $x = T$, $y = T$, and $z = T$. Since,

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

$$\begin{aligned} f(x = T, y = T, z = T) &= (T \circ T \circ T) \bullet (a \circ F \circ T \circ T) \\ &= T \bullet F \\ &= T \end{aligned}$$

the functional assignment is T and the updated cell color is Black. Therefore, the very first digit in the list must be a 2.

The 23rd arrangement of colors at time step N is WRR. This corresponds to the arguments $x = F$, $y = a$, and $z = a$. Since,

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

$$\begin{aligned} f(x = F, y = a, z = a) &= (F \circ a \circ a) \bullet (a \circ T \circ a \circ a) \\ &= F \bullet a \\ &= a \end{aligned}$$

the functional assignment is a and the updated cell color is Red. Therefore, the 23rd digit in the list must be a 1.

When all is said and done, the effort expended in creating the list of 27 digits in this manner is submitted to **FromDigits[ ]** to find the appropriate rule number for a three color CA,

```
FromDigits[{2, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0,
                                  1, 1, 0, 1, 1, 0, 0, 0, 0}, 3]
```

which returned the rule number 6 056 723 077 785.

After running,

```
RulePlot[CellularAutomaton[{6 056 723 077 785, 3, 1}],
                 ColorRules → {0 → White, 1 → Red, 2 → Black}]
```

it is easy to confirm that the proper updated cell colors are associated with all possible three color arrangements at the previous time step.

Supplemental Exercise 3.3.5: Figure out the color for any updated cell in a four color CA.

Solution to Supplemental Exercise 3.3.5:

We know that there will be an even more enormous number of possible four color CA than were envisioned above for any three color CA. There are $4^{4^3} \approx 3.4 \times 10^{38}$ possible CA.

We could select 0 or any integer up to the maximum integer consisting of 39 digits. I arbitrarily selected rule number 45 609 287 653 895 678 consisting of only 17 digits as an argument for **IntegerDigits[ ]**. There will now be a total of $4^3 = 64$ possible triples of cell colors that need an assignment from say, yellow, blue, red, and now an additional cell color, say, green.

Evaluating the rule number as a base 4 number with 64 possible terms,

```
IntegerDigits[45 609 287 653 895 678, 4, 64]
```

returns a long list of 64 elements consisting of 0s, 1s, 2s and 3s. The list starts with 36 0s and ends with 28 non–zeros, **{0, 0, ···, 3, 1, 3, 3, 3, 2}**.

For the sake of the example, select the 3, the sixth value over from the right in the list. This updated cell must be colored green because of the 3. What are the colors of the three cells that this rule says will update to a green cell? Start at YYY and rotate through to YYB, YYR, YYG, YBY, and finally to YBB. We now know that when the above three relevant cells are colored Yellow, Blue, and Blue, the cell will be updated to the color green.

Supplemental Exercise 3.3.6: Make a conjecture about the relative sizes of all possible CA and all possible Boolean functions.

Solution to Supplemental Exercise 3.3.6:

I conjecture that the set of all possible CA is much larger than the corresponding set of all Boolean functions when $k > 2$. What particular example was I working on that led me to make this conjecture?

I happened to be working on the same example as in Exercise 3.3.2 for the CA following rule 2 625 597 484 982. I mistakenly thought that every such CA would have its counterpart as some Boolean function. But when I constructed the truth table based on this CA, the resulting MCF expansion of the Boolean function would not work.

The truth table will consist of $2^n = 2^3 = 8$ rows starting out at the variable assignment of T, T, T and ending with F, F, F. Remember that the three colors used were yellow (Y), blue (B), and red (R) corrresponding to F, $a = a'$, and T. See Table 3.2 below.

Table 3.2: *Truth table for some unknown Boolean function.*

Row	xyz	$f(x, y, z)$	*Previous Colors*	*Updated Color*
1	TTT	a	RRR	B
2	TTF	F	RRY	Y
3	TFT	F	RYR	Y
4	TFF	F	RYY	Y
5	FTT	T	YRR	R
6	FTF	T	YRY	R
7	FFT	a	YYR	B
8	FFF	T	YYY	R

The minterm canonical form derived from this truth table is,

$$f(x, y, z) = axyz \bullet x'yz \bullet x'yz' \bullet ax'y'z \bullet x'y'z'$$

Consult the rule plot for the CA to verify any variable assignment for this Boolean function. For example, the very first assignment,

$$f(x = T, y = T, z = T) = (a \circ T \circ T \circ T) \bullet F \bullet F \bullet F \bullet F = a \equiv \text{ BLUE}$$

which is correct. Notice that the last four terms all have an x' making them all F.

The next assignment from the Boolean function,

$$f(x = T, y = T, z = a) = (a \circ T \circ T \circ a) \bullet F \bullet F \bullet F \bullet F = a \equiv \text{ BLUE}$$

also indicates a BLUE updated cell because of the first term $a \circ T \circ T \circ a$ with the remaining four terms all F for the same reason as above. However, the cell is actually colored YELLOW according to the rule plot.

To confirm this, look at the second digit from the left resulting from our initial examination of **`IntegerDigits[2 625 597 484 982, 3, 27]`**. It is a 0 indicating a YELLOW cell for the previous cell colors of RRB. The coefficient for the building block function 3^{25} in the decimal expansion is 0, just as the coefficient for the building block function 3^{26} in the decimal expansion was 1.

For the final straw to break the camel's back, look at the next to the final assignment,

$$f(x = F, y = F, z = a) = a \equiv \text{ BLUE}$$

The first three terms are all F, the fourth is a, and the fifth is $a' = a$, and all of this leading to $F \bullet F \bullet F \bullet a \bullet a = a$. But according to **`RulePlot[ ]`**, YYB does NOT result in a BLUE cell but instead a RED cell.

The final lesson is this: Even though every specified Boolean function can be turned into an analogous CA, not every specified CA with cell colors greater than two can be turned into an analogous Boolean function. These CA are some form of a mapping, but they are not Boolean functions.

Supplemental Exercise 3.3.7: Solve Brown's Exercise 4 in Chapter 3 for further insight on this issue.

Solution to Supplemental Exercise 3.3.7:

My eyes were opened about the necessary distinction in defining a *Boolean function* when I came across this exercise [5, pg. 71, Exercise 4]. Brown sets up his carrier set as,

$$\mathbf{B} = \{0, 1, a', a\}$$

He then asks us to determine which of three single variable functions is a Boolean function. It turns out that $g(x)$ is a Boolean function, but $f(x)$ and $h(x)$ are not. He shows us the functional assignment table for all three functions as given in Table 3.3 below.

Table 3.3: *Functional assignment table for three functions.*

x	$f(x)$	$g(x)$	$h(x)$
0	a	a'	a'
1	a	1	1
a'	a'	a'	1
a	a'	1	a'

The minterm canonical form will be very simple for a one variable Boolean function. In my notation, the MCF would be,

$$\mathcal{F}(x) = \mathcal{F}(T)\, x \bullet \mathcal{F}(F)\, x'$$

The MCF for the three functions are then,

$$f(x) = ax \bullet ax'$$

$$g(x) = x \bullet a'x'$$

$$h(x) = x \bullet a'x'$$

If $f(x)$ were a Boolean function, then,

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

So, $f(x)$ is a constant function with the consequence that $f(x = a')$ and $f(x = a)$ should both equal a which they do not. If $h(x)$ were a Boolean function, then,

$$h(x = a') = a' \bullet (a' \circ a) = a' \bullet F = a'$$

but T is given as the functional assignment ruling $h(x)$ out as a Boolean function.

If $g(x)$ is to be a Boolean function, then we must check that all four functional assignments from Table 3.3 are correct,

$$g(x = F) = x \bullet a'x' = F \bullet (a' \circ T) = F \bullet a' = a'$$

$$g(x = T) = x \bullet a'x' = T \bullet (a' \circ F) = T \bullet F = T$$

$$g(x = a') = x \bullet a'x' = a' \bullet (a' \circ a) = a' \bullet F = a'$$

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

which they are.

Chapter 4

Analogies Between Formal Manipulations

4.1 Ordering Relations in Boolean Algebra

The whole idea of "ordering relationships" in a Boolean Algebra seemed to come through in a less transparent manner than I had intended for Chapter Four of Volume I. After some further review of these concepts as they appear formally within Boolean Algebra, we continue to emphasize that the primary motivation is to investigate the analogies with probability assignments.

Let the symbols a and b stand for two elements of a carrier set $\mathbf{B}$. Then, an ordering relation, "a is less than or equal to b" is defined for a and b through,

$$a \leq b \equiv a \circ b' = F$$

This formal definition of an ordering within Boolean Algebra has its parallel in probability assignments when these assignments are conditioned on the information from models inspired by the forward implication logic function. Thus,

$$P(A\overline{B} \,|\, \mathcal{M}_k \equiv A \rightarrow B) = 0$$

Supplemental Exercise 4.1.1: Revisit the expansion for the implication logic function.

Solution to Supplemental Exercise 4.1.1:

Boole's Expansion Theorem for the forward implication logic function is familiar by now. $f_{13}(A, B) \equiv A \rightarrow B$ follows the standard template,

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

When the coefficients for this particular function are defined by,

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

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

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

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

the minterm canonical form, or alternatively, the fully expanded DNF for $f_{13}(A,B)$ becomes $MCF\,[\,f_{13}(A,B)\,] = AB \vee \overline{A}B \vee \overline{A}\,\overline{B}$. Therefore, the ordering relation of "less than or equal," $a \circ b' = F$, is analogous to a model inspired by the forward implication logic function, and its implementation in probability as an assignment of $P(A\overline{B}\,|\,\mathcal{M}_k \equiv A \rightarrow B) = 0$.

Supplemental Exercise 4.1.2: Draw out some further properties in a Boolean Algebra given an ordering relationship between two elements.

Solution to Supplemental Exercise 4.1.2:

Assert equivalence between $a \leq b$ and $a \circ b' = F$. Then, invoke the **Duality Principle**,

$$a \circ b' = F \overset{\text{Duality}}{\Longrightarrow} a' \bullet b = T$$

Substituting $b = a$, we have $a' \bullet a = T$ and $P(\overline{A}) + P(A) = 1$.

Retaining the more general $a' \bullet b = T$, the analogous probability assignment is,

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

This is easily seen to be correct after expanding each of the two terms on the left hand side,

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

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

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

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

And finishing up with,

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

$$P(A\overline{B}) = 0$$

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

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

Of course, we have always recommended that formal manipulations like the above be viewed in the context of a joint probability table. When the four joint statements are set up as the four cells of a joint probability table, and with cell 3, $P(A\overline{B}) = 0$, it is obvious that cells 1, 2, and 4 must add up to 1.

Furthermore, $P(\overline{A})$ is represented by the sum of the assignments in cells 2 and 4, and $P(B)$ is represented by the sum of the assignments in cells 1 and 2. Subtracting off the repeated assignment in cell 2 by the formal manipulation rule,

$$\begin{aligned}
P(\overline{A} \vee B) &= P(\overline{A}) + P(B) - P(\overline{A}B) \\
&= \text{cell 2} + \text{cell 4} + \text{cell 1} + \text{cell 2} - \text{cell 2} \\
&= \text{cell 1} + \text{cell 2} + \text{cell 4} \\
&= P(AB) + P(\overline{A}B) + P(\overline{A}\,\overline{B}) \\
P(A\overline{B}) &= 0 \\
P(\overline{A} \vee B) &= 1
\end{aligned}$$

Supplemental Exercise 4.1.3: In a reverse fashion, use the argument from the joint probability table to assert that in a Boolean Algebra the ordering relationship imposes $a \bullet b = b$.

Solution to Supplemental Exercise 4.1.3:

As in the above development, imposition of the ordering relationship $a \leq b$ from Boolean Algebra has the probabilistic consequence that $P(A\overline{B}) = 0$. $P(A \vee B) = P(A) + P(B) - P(AB)$ is cell 1 + cell 3 + cell 1 + cell 2 − cell 1 with cell 3 = 0. Then, $P(A \vee B) = \text{cell 1} + \text{cell 2} = P(B) \equiv a \bullet b = b$.

Supplemental Exercise 4.1.4: In Exercise 4.6.3 of Volume I, as part of the task of filling out the operator table for $\circ$, it was claimed that $a' \circ b' = b'$. Prove this by arguing from the perspective of a joint probability table.

Solution to Supplemental Exercise 4.1.4:

From the **Duality Principle**, given the correctness of $a \bullet b = b$ from the last exercise, we have immediately that $a' \circ b' = b'$. Once again, from the ordering relationship, $P(A\overline{B}) = 0$. It is easy to discern from the joint probability table that $P(\overline{B}) = P(A\overline{B}) + P(\overline{A}\,\overline{B})$, that is, the marginal probability for the statement "B is FALSE." is the probability for the statement, "A and B are both FALSE."

Supplemental Exercise 4.1.5: In Exercise 4.6.4 of Volume I, as part of the task of filling out the operator table for $\bullet$, it was claimed that $a' \bullet b' = a'$. Prove this.

Solution to Supplemental Exercise 4.1.5:

Arguing from the perspective of a joint probability table with $P(A\overline{B}) = 0$,

$$\begin{aligned} P(\overline{A} \vee \overline{B}) &= P(\overline{A}) + P(\overline{B}) - P(\overline{A}\,\overline{B}) \\ &= P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) - P(\overline{A}\,\overline{B}) \\ &= P(\overline{A}B) + P(\overline{A}\,\overline{B}) + P(A\overline{B}) \\ &= P(\overline{A}B) + P(\overline{A}\,\overline{B}) \\ &= P(\overline{A}) \end{aligned}$$

Supplemental Exercise 4.1.6: Establish an ordering relationship among the elements of a carrier set. Illustrate the relationship with a "Venn diagram."

Solution to Supplemental Exercise 4.1.6:

Suppose that the elements of a carrier set for a Boolean Algebra consist of,

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

If $a \leq b$, then we have $F \leq a \leq b \leq b' \leq a' \leq T$. Figure 4.1 below shows a Venn diagram for this ordering relationship. Such a diagram may make it more plausible for identities like $a \bullet b = b$ and $a' \bullet b' = a'$.

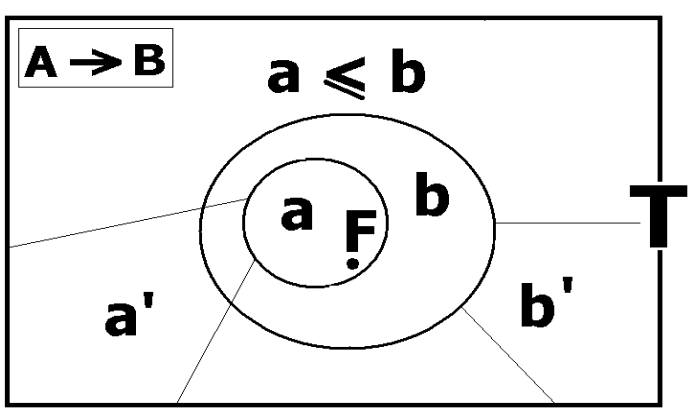

Figure 4.1: *A Venn diagram illustrating an ordering relationship.*

I must mention that I don't particularly like using Venn diagrams to illustrate these ordering relationships because none of F, a, a', b, b', or T are, in fact, sets containing a number of elements. Nonetheless, since so many people are accustomed to viewing set relationships from a Venn diagram, I draw upon the visual analogy of set theory conflated with Boolean Algebra.

Even though I pictured the "set" F as a small circle sitting inside all of the other circles, it should be thought of as infinitesimally small with no other set being able to fit inside it. Likewise, all of the other sets are able to fit inside the "universal set" T. The sectors marked out for a' and a show just a portion of the sets.

4.2 Analogies with Probability Assignments

All of these abstract concepts involving a Boolean Algebra assume a rather mundane interpretation within probability theory. The joint probability table is the vehicle that I choose to exhibit a transparent instantiation of these notions. By far the most important concept, and, furthermore, the one easiest to grasp is that it is the information in a model inspired by the implication logic function that leads to the same relationships in probability theory as asserted by the ordering relationship in Boolean Algebra.

Supplemental Exercise 4.2.1: What is the minterm canonical form for the Boolean function of two variables implementing the ordering relationship $a \leq b$?

Solution to Supplemental Exercise 4.2.1:

Once again, invoking the generic template for the expansion of a two variable Boolean function with its four constituent building block functions and associated coefficients would look like,

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

The four coefficients are determined by the operator table of Table 4.4, Volume I, as,

$$f(T,T) = a \circ b = a$$

$$f(T,F) = a \circ b' = F$$

$$f(F,T) = a' \circ b = b$$

$$f(F,F) = a' \circ b' = b'$$

The minterm canonical form is then,

$$MCF\,[\,f(x,y)\,] = axy \bullet bx'y \bullet b'x'y'$$

Supplemental Exercise 4.2.2: Demonstrate that the above expression doesn't seem to work out for the typical 2×2 joint probability table implementing the implication function.

Solution to Supplemental Exercise 4.2.2:

If we were to make the associations from the elements in the carrier set to probability assignments in a four cell joint probability table as shown in Figure 4.2 below, there is something wrong. Cell 2, $a' \circ b = b$, should equal 1/3, but $b = P(B) = 2/3$. There are other discrepancies as well. Cell 2 becomes $a' \circ b = c'$ as shown in the next exercise.

	A	$\overline{A}$	
B	Cell 1 1/3 $a \circ b = a$	Cell 2 1/3 $a' \circ b = b$	2/3
$\overline{B}$	Cell 3 0 $a \circ b' = F$	Cell 4 1/3 $a' \circ b' = b'$	1/3
	1/3	2/3	1

Figure 4.2: *A four cell joint probability table with numerical assignments inspired by the forward implication logic function.*

Supplemental Exercise 4.2.3: How was this corrected?

Solution to Supplemental Exercise 4.2.3:

The carrier set was enlarged to include new elements c and c'. One new coefficient was defined as c', to replace the assignment of b, while the other three coefficients remained the same,

$$f(T,T) = a \circ b = a$$

$$f(T,F) = a \circ b' = F$$

$$f(F,T) = a' \circ b = c'$$

$$f(F,F) = a' \circ b' = b'$$

leading to the new minterm canonical form of the function as shown on page 85, Volume I,

$$MCF\,[\,f(x,y)\,] = axy \bullet c'x'y \bullet b'x'y'$$

For example, cell 1 of the joint probability table, where $x = T$ and $y = T$, the function returns a value of a analogous to the probability assignment in cell 1 under the information from a model inspired by the forward implication logic function,

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

Supplemental Exercise 4.2.4: What does the functional value at cell 4 of the joint probability table tell us?

Solution to Supplemental Exercise 4.2.4:

The functional value corresponding to cell 4 is,

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

If $a' \circ b' = b'$, then by the **Duality Principle**,

$$a' \circ b' = b' \overset{\text{Duality}}{\Longrightarrow} a \bullet b = b$$

Does $P(A \vee B) = P(B)$? Yes, as was shown in Supplemental Exercise 4.1.3,

$$P(A \vee B) = P(A) + P(B) - P(AB) = 1/3 + 2/3 - 1/3 = 2/3 = P(B)$$

Chapter 5

Fundamental Rules of Probability

We resort to the helpful analogies borrowed from Boolean Algebra, as begun in Chapter Four of Volume I, while transitioning to the formal rules for probability manipulations.

In defense of my non–rigorous arguments shuttling back and forth between Boolean operations and probability operations, I appeal to the sometimes fruitful approach used by mathematicians of "plausible conjectures through analogy."

More often than not, one is blind to seeing the next steps that must be taken to move forward. Generating plausible conjectures when stymied is one way in which better explanations may be formed during the creation of new knowledge.

The following is an apposite quote from Jaynes [11, pg. 6] appearing in his book expressing the same sympathies.

> Pólya showed that even a pure mathematician actually uses these weaker forms of reasoning most of the time. Of course, on publishing a new theorem, the mathematician will try very hard to invent an argument which uses only the first kind [deductive reasoning]; but the reasoning process which led to the theorem in the first place almost always involves one of the weaker forms (based, for example, on following up conjectures suggested by analogies). The same idea is expressed in a remark of S. Banach (quoted by S. Ulam, 1957):
>
>> Good mathematicians see analogies between theorems; great mathematicians see analogies between analogies.

So, once again, I am echoing a common refrain that starts with Boolean Algebra, continues on into Classical Logic, and then runs through to the generalizations inherent in Probability Theory.

5.1 Boolean Operations and Probability

Supplemental Exercise 5.1.1: Verify that the entries in all four cells of the joint probability table of Figure 5.1 sum to T.

Solution to Supplemental Exercise 5.1.1:

Each entry in the joint probability table is an element from the carrier set of a Boolean Algebra instead of a some probability assignment. This entry into each cell of the joint probability table results from applying the $\circ$ and $\bullet$ Boolean operators as given in Tables 4.6 and 4.7 of Volume I. Apply the **Associativity axiom** to illustrate that, however the entries might be grouped by parentheses, the answer is always T. Figure 5.1 below is an enhanced version of the joint probability table that also appears as Figure 5.1 in Volume I.

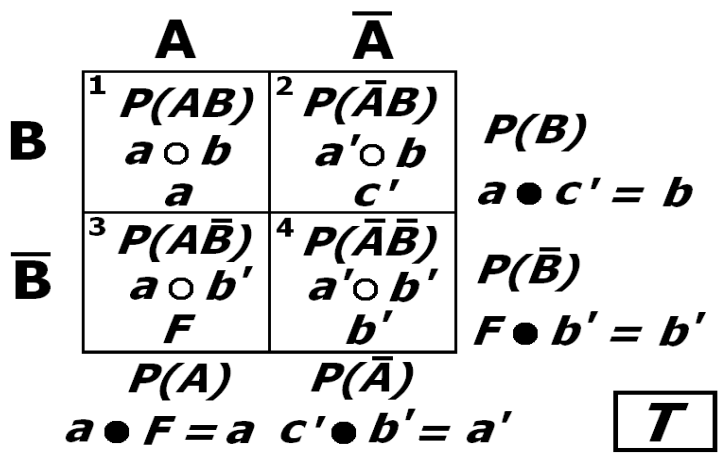

Figure 5.1: *A symbolic joint probability table using Boolean operations.*

The results for each cell may be found by consulting Volume I's Table 4.6 for the binary $\circ$ operator,

$$a \circ b = a$$

$$a' \circ b = c'$$

$$a \circ b' = F$$

$$a' \circ b' = b'$$

The Boolean Algebra ordering relationship of $a \leq b$ is in effect since $a \circ b' = F$. Apply the binary $\bullet$ operator to these four entries by consulting Volume I's Table 4.7 to find out if,

$$a \bullet c' \bullet F \bullet b' = T$$

Parentheses must used in the expression to indicate the order in which the operations are to take place. Examine the particular starting order of,

$$a \bullet ((c' \bullet F) \bullet b')$$

to find that the result does equal T,

$$a \bullet ((c' \bullet F) \bullet b') \rightarrow a \bullet (c' \bullet b) \rightarrow a \bullet a' = T$$

or, in terms of summing the probabilities over the two rows in the joint probability table, $P(A \vee \overline{A}) = P(A) + P(\overline{A}) = 1$.

Examine a different starting order by placing the parentheses as,

$$((a \bullet c') \bullet F) \bullet b')$$

to find that the result once again does equal T,

$$((a \bullet c') \bullet F) \bullet b') \rightarrow (b \bullet F) \bullet b' \rightarrow b \bullet b' = T$$

Again, in terms of summing the probabilities over the two columns in the joint probability table, $P(B \vee \overline{B}) = P(B) + P(\overline{B}) = 1$.

Recall that it is the information in the implicational model $A \rightarrow B$ that provides the justification for the entries in the cells of the joint probability table as numerical values of probabilities. The ordering relationship in a Boolean Algebra of $a \leq b$ provides the analogous symbolic entries from the carrier set.

Supplemental Exercise 5.1.2: Show the analogous Boolean operations for the derivation of the distribution of the probability operator at the end of section 5.5 of Volume I.

Solution to Supplemental Exercise 5.1.2:

Step 1. De Morgan's axiom $(x' \circ y') = (x \bullet y)' \Longrightarrow P(\overline{A}\,\overline{B}) = P(\overline{A \vee B})$

Step 2. Complementation $x' \bullet x = T \Longrightarrow P(\overline{\overline{A \vee B}}) + P(\overline{A \vee B}) = 1$

Step 3. Involution $(x')' = x \Longrightarrow P(\overline{\overline{A \vee B}}) = P(A \vee B)$

Work backwards from the probabilistic answer of $P(A \vee B) + P(\overline{A \vee B}) = 1$ to the equivalent answer in Boolean Algebra,

$$(x \bullet y) \bullet (x \bullet y)' = (x \bullet y) \bullet (x' \circ y') \qquad \textbf{De Morgan}$$

$$(x \bullet y) \bullet (x' \circ y') = (x \bullet y) \bullet (y' \circ x') \qquad \textbf{Commutativity}$$

$$(x \bullet y) \bullet (y' \circ x') = x \bullet ((y \bullet y') \circ x') \qquad \textbf{Associativity}$$

$$x \bullet ((y \bullet y') \circ x') = x \bullet (T \circ x') \qquad \textbf{Complementation}$$

$$x \bullet (T \circ x') = x \bullet x' \qquad \textbf{Special Elements}$$

$$x \bullet x' = T \qquad \textbf{Complementation}$$

Supplemental Exercise 5.1.3: Engage in another interesting exercise in manipulating the probability expressions that naturally arise during the proof of distributing probability over the $\vee$ symbol.

Solution to Supplemental Exercise 5.1.3:

This exercise was inspired by a derivation of another proof that,

$$P(A \vee B) = P(A) + P(B) - P(AB)$$

as given by Anthony Garrett [7]. For a little variety, I won't slavishly duplicate the forward order of steps in Garrett's proof. I will instead reverse the process and start with the expression that was to be proven and proceed backwards.

To get us in the right mood, recall **De Morgan's axiom** from Boolean Algebra, and its translation over into the analogous probability expression.

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

$$1 - P(A \vee B) = P(\overline{A}\,\overline{B})$$

$$P(A \vee B) = 1 - P(\overline{A}\,\overline{B})$$

The following series of steps take us from $P(A) + P(B) - P(AB)$ to $P(A \vee B)$ through a recognized transformation permitted by the formal manipulation rules of probability at each step.

$$\begin{aligned}
P(A) + P(B) - P(AB) &= P(B) + P(A) - P(B \,|\, A\,)\, P(A) \\
&= P(B) + [\,1 - P(B \,|\, A)\,]\, P(A) \\
&= P(B) + P(\overline{B} \,|\, A)\, P(A) \\
&= [\,1 - P(\overline{B})\,] + P(\overline{B}A) \\
&= 1 - P(\overline{B}) + P(A\overline{B}) \\
P(A\overline{B}) &= [\,1 - P(\overline{A} \,|\, B\,]\, P(\overline{B}) \\
P(A) + P(B) - P(AB) &= 1 - P(\overline{B}) + [\,1 - P(\overline{A} \,|\, \overline{B})\,]\, P(\overline{B}) \\
&= 1 - P(\overline{B}) + P(\overline{B}) - P(\overline{A} \,|\, \overline{B})\, P(\overline{B}) \\
&= 1 - P(\overline{A}\,\overline{B}) \\
&= P(A \vee B)
\end{aligned}$$

Supplemental Exercise 5.1.4: Prove that the degree of belief in the truth that a statement is both TRUE and FALSE is zero.

Solution to Supplemental Exercise 5.1.4:

In Exercise 5.9.1 of Volume I, some basic manipulation rules of probability were listed, one of which was $P(A\overline{A}) = 0$. Begin with the Boolean axiom $x \bullet x' = T$. Through the duality principle, by which we mean invoking **De Morgan's axiom**, **Complementation**, and **Commutativity** we have,

$$x \bullet x' = T \Longrightarrow (x' \circ x) = T' \Longrightarrow x' \circ x = F \Longrightarrow x \circ x' = F$$

leading to the probabilistic assertion, by analogy with the above Boolean operations, that $P(A \wedge \overline{A}) \equiv P(A\overline{A}) = 0$.

Supplemental Exercise 5.1.5: Show another Boolean inspired probability relationship involving three statements.

Solution to Supplemental Exercise 5.1.5:

This exercise is a variation on the probability relationships derived in Exercises 5.9.9, 5.9.12, and 5.9.13 of Volume I.

From $x \bullet x' = T$,

$$P(ABC) + P(\overline{ABC}) = 1$$

From $((x \circ y) \circ z)' = (x' \bullet y') \bullet z'$ and the above,

$$P(ABC) + P(\overline{A} \vee \overline{B} \vee \overline{C}) = 1$$

$$P(ABC) = 1 - P(\overline{A} \vee \overline{B} \vee \overline{C})$$

Since $P(ABC)$ is located in cell 1 of the joint probability table, $P(\overline{A} \vee \overline{B} \vee \overline{C})$ must be the sum of the probabilities located in cells 2 through 8.

Supplemental Exercise 5.1.6: Given some information provided by a model, what is the probability of the joint statement *ABC*?

Solution to Supplemental Exercise 5.1.6:

It depends on what information is guiding the probability assignments to the joint statements in the joint probability table. Suppose that Rule 129 of the elementary cellular automata is an overarching model for these assignments. Then, the answer is that the probability for the joint statement ABC is one minus the probability assigned to cell 8 of the joint probability table.

The binary number representing Rule 129 is 1000001. Matching the 1s and 0s in this binary number to all eight possible variable settings, we see that the DNF for the Boolean function representing Rule 129 is $ABC \vee \overline{A}\,\overline{B}\,\overline{C}$. The leftmost 1 matches up with TTT and the rightmost 1 matches up with FFF. The 0s match up with the other six possibilities. The coefficient for the building block functions $A \wedge B \wedge C$ and $\overline{A} \wedge \overline{B} \wedge \overline{C}$ in the orthonormal function expansion is T, while the coefficients for the other six building block functions are all F.

From the last exercise, we have,

$$P(ABC) = 1 - P(\overline{A} \vee \overline{B} \vee \overline{C})$$

But the assignments to cells 2 though 7 must be 0 given the assignments made under the information provided by Rule 129.

There must be some legitimate probability assignment to both cell 1 and cell 8. Therefore, $P(ABC) = 1 - P(\overline{A}\,\overline{B}\,\overline{C})$, or 1 minus the probability assigned to cell 8 of the joint probability table. For example, $P(ABC) = 1/3$ and $P(\overline{A}\,\overline{B}\,\overline{C}) = 2/3$ with the other six cells of the joint probability table assigned 0s. Neither cell 1 nor cell 8 can be assigned a 0 under the model of Rule 129.

5.2 Conceptual Foundations

A great debt is owed to Sir Harold Jeffreys for laying out, in the clearest possible fashion, (except for his confusing notation), the conceptual foundations behind the formal manipulation rules as they are universally employed in probability.

I recommend first reading his Chapter 2 of **Scientific Inference** [13] before dipping into his **Theory of Probability** for an initial exposure to these formal rules. I think one can justifiably argue that his first book, especially the sections concerning the foundations for probability, are the product of Jeffreys's thinking covering roughly a twenty year period from about 1915 through 1935.

I have closely followed Jeffreys's axioms in presenting my own introduction. Of course, this is not saying all that much. Many people, especially those who are of the Bayesian stripe, have taken Jeffreys as sacred writ. Jaynes for one acknowledged that it was reading Jeffreys as a Ph. D. candidate in Physics at Princeton in the 1940s that opened his eyes to the proper way to conduct scientific inference.

The modern reader does experience some difficulty with Jeffreys's writing style. He is, in my opinion, a classic example of the identifiable style of a British don writing for an early 20th Century audience. I doubt that there would be much disagreement that Jeffreys, as well as his intended audience, were products of an elite, upper class Victorian educational system.

One hallmark of such an immersion is a tone of *reserved understatement*. As a consequence, ideas of profound importance are often tossed your way unburdened by any noticeable passion or enthusiasm. One must always maintain a proper British gentlemen's reserve; especially when contrasted to a typical American's unrestrained excitement.

As an introductory example of this style, and really one of the most profound statements underlying the conceptual foundation of probability theory is,

> If we like there is no harm in saying that a probability expresses a degree of reasonable belief.

We do hope that no harm follows from holding this idea close to our hearts. Personally, I phrase it this way: From an information processing standpoint, a probability express a quantitative degree of belief as held by an IP in the truth of some statement under the information provided by some model.

Implicit in Jeffreys's laconic characterization of probability is the core idea that probability must be an epistemological notion. Probabilities may exist only in the conscious mind of an IP. Probabilities are not to be found in any ontological property of the world the IP inhabits, quantum mechanics notwithstanding.

I think that Jaynes assimilated, and consequently actively promoted the idea that probability theory was indeed a generalization of Classical Logic after reading Jeffreys. To wit,

[13], p. 23

[13] Jeffreys, Harold. Scientific Inference, Second Edition, Cambridge University Press, 1957.

> When we make an inference beyond the observational data, we express a logical relation between the data and the inference. This relation is in a generalized logic, not in deductive logic. It does not claim that the inference is deductively proved or disproved from the data. It assesses the support for the inference, given the data, but an essential feature is that this support can be of many different degrees.

The main difficulty in reading Jeffreys is in his choice of notation. He uses the symbols p, q, r, and so on, as borrowed from the abstract propositions of logic; we prefer notation like A, B, C, and so on to indicate *statements*. What is more objectionable is his constant insistence on language that labels p as "data." This is very confusing to the reader.

> In probability theory both the data and the proposition considered are subject to alteration, and it is therefore necessary to keep the data explicit. This is achieved by writing the relation in the form
>
> $$P(q \mid p) = a$$
>
> (read 'the probability of q, given p, is a'), where a is the number that expresses the degree of confirmation.

As was obvious from Chapter Five of Volume I, the mutually exclusive and exhaustive properties of joint statements as indexed by the cells of joint probability

tables were viewed as a superior way to justify basic probability manipulations. Jeffreys did not choose to argue in this manner, but rather gave more difficult justifications that demand a lot more focused attention on the part of the reader. My hope is that after absorbing my style of justifying probability manipulations as presented in Chapter Five, any one will have an easier experience absorbing the important content of Jeffreys's Chapter 2.

Supplemental Exercise 5.2.1: How do we write and verbally translate Jeffreys's probability?

Solution to Supplemental Exercise 5.2.1:

Instead of Jeffreys's $P(q\,|\,p) = a$, we write,

$$P(A = a_i\,|\,\mathcal{M}_k) = Q_i$$

We do retain Jeffreys's notation of an uppercase P to represent a probability, the parentheses, and also the *solidus*, that is, the "conditioned upon" symbol "|" within a general notational template of $P(\star\,|\,\star)$ where $\star$ must stand for some statement.

Our probability expression translated into words goes something like this: "*One* legitimate number assigned as *a* probability to some statement $(A = a_i)$ conditioned on the truth of the information inserted by the statement of model $\mathcal{M}_k$." We prefer not to use language like "*the* probability *of* q" because it tends to convey the impression that probability represents a physical property possessed by some object.

The symbol a is replaced by Q_i, but of course must remain, as Jeffreys says, a real number between 0 and 1 inclusive. The symbol p as replaced by $\mathcal{M}_k$ is not *data* in any sense of the word. *Data* are the collection of the observed occurrences of any of the statements in the state space, $\mathcal{D} = \{A_1 = a_i, A_2 = a_j, \cdots, A_N = a_k\}$

Supplemental Exercise 5.2.2: What two fundamental manipulation rules did Jeffreys present next?

Solution to Supplemental Exercise 5.2.2:

After introducing a few axioms, Jeffreys gets around to presenting the **Sum Rule** and **Product Rule**. These two formal manipulation rules are in constant use as we saw beginning in Chapter Five. Jeffreys writes the **Sum Rule** as,

$$P(q\,|\,p) = P(q.r\,|\,p) + P(q. \sim r\,|\,p) \geq P(q.r\,|\,p)$$

Jeffreys uses the symbol "." for the `And` between two propositions, and the symbol "$\sim$" for negation. We use $\wedge$ and $\vdash$. As always with Jeffreys, these symbolic

expressions represent probabilities conditioned on "data p." We write instead,

$$\begin{aligned} P(B \,|\, \mathcal{M}_k) &= P(BA \,|\, \mathcal{M}_k) + P(B\overline{A} \,|\, \mathcal{M}_k) \geq P(BA \,|\, \mathcal{M}_k) \\ &= P(AB \,|\, \mathcal{M}_k) + P(\overline{A}B \,|\, \mathcal{M}_k) \geq P(AB \,|\, \mathcal{M}_k) \\ &= P(B \,|\, \mathcal{M}_k) \geq P(AB \,|\, \mathcal{M}_k) \end{aligned}$$

because we are projecting ahead to the use of the **Sum Rule** in Bayes's Theorem,

$$\begin{aligned} P(A \,|\, B, \mathcal{M}_k) &= \frac{P(AB \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \\ &= \frac{P(AB \,|\, \mathcal{M}_k)}{P(AB \,|\, \mathcal{M}_k) + P(\overline{A}B \,|\, \mathcal{M}_k)} \end{aligned}$$

The **Sum Rule** is supported through the analogy with the Boolean operations,

$$\begin{aligned} P(BA) + P(B\overline{A}) &\Longrightarrow (B \wedge A) \vee (B \wedge \overline{A}) \\ &= B \wedge (A \vee \overline{A}) \\ &= B \wedge T \\ &= B \end{aligned}$$

To reaffirm Jeffreys's meaning for the symbol p, look at his preface to his first axiom.

> Our fundamental hypothesis is that degrees of confirmation of different propositions, on the same data, can be put in order, so that our fundamental type of comparison is of the form 'on data p, q is more probable than r'.

If not obvious by now, any time Jeffreys says "conditioned on data p" replace this with "conditioned on assuming the truth of the information inserted into a probability distribution by the model $\mathcal{M}_k$."

Jeffreys presents the **Product Rule** written as,

$$P(q.r \,|\, p) = P(q \,|\, p)\, P(r \,|\, q.p)$$

while we prefer,

$$P(A, B \,|\, \mathcal{M}_k) = P(A \,|\, B, \mathcal{M}_k)\, P(B \,|\, \mathcal{M}_k)$$

Supplemental Exercise 5.2.3: What pitfall did Jeffreys warn us about?

Solution to Supplemental Exercise 5.2.3:

As we shall examine in more detail in the upcoming Chapters Seven and Ten of these **Supplemental Exercises**, Jeffreys provides a warning when engaging in

probability manipulations. Basically, the warning is to carefully examine what inference you want to make if it should ever happen that the denominator in Bayes's Theorem should equal zero.

> Thus all our axioms need elaboration by a statement that the data must not be self–contradictory ... In science we are not interested in inferences from self–contradictory data; a contradiction among our data, if there is one, would be looked for by deductive methods and the data would be modified accordingly. (There could of course be no contradiction among observational data, but there might be one between two hypotheses or between observational data and some hypothesis whose consequences are being examined.)

This does NOT mean that a statement that is assumed FALSE may not appear to the right of the conditioned upon symbol. It means that any statement that appears to the right of the conditioned upon symbol, together with the model $\mathcal{M}_k$ that appears to the right may not be contradictory. Several examples are upcoming.

For me, Jeffreys's warning is also suggestive of something else. This may seem like rather unmotivated advice at this stage, but I think its meaning will become quite clear later. Jeffreys's warning advises an IP not to fixate on any one single model; it should carry along all possible models, and let the rules of probability theory winnow out the wheat from the chaff.

Supplemental Exercise 5.2.4: Why is Jeffreys held in such high esteem by Bayesians?

Solution to Supplemental Exercise 5.2.4:

After carefully laying out these fundamental considerations, and formal rules for probability manipulations, Jeffreys launches into what must be judged as his *pièce de résistance*. He argues that the next plausible, and culminating step, in the development in these fundamental axioms leads to Bayes's Theorem.

Although he didn't say this explicitly, (although I think he should have), some language along these lines would have been welcome: "After all of this build–up, I am now going to show you how to generalize logic and, even more importantly, provide you with the overarching framework for how to conduct scientific inference."

Remember though, to an Englishman of the right type, reserved understatements are always to be preferred! He begins with,

> Suppose that there is some general datum that is included in all experience; it might be for instance be the rules of pure mathematics. Denote this by H. If proposition p has a positive (non–zero) probability on data H and q is any other proposition, assume that $P(q\,|\,p.H)$ satisfies
>
> $$P(q\,|\,p.H) = P(p.q\,|\,H)/P(p\,|\,H)$$

I don't believe it was necessary to introduce the new symbol H to represent everything we might be assuming as true. The information in model $\mathcal{M}_k$ continues to serve this role without modification. And it changes p from how Jeffreys had formerly defined it into now just another proposition like q. In addition, notice that there is no mention by Jeffreys of this relationship he just spelled out as being something called *Bayes's Theorem*.

By now, you know how Jeffreys's notation gets translated into mine. The basic generic expression for Bayes's Theorem that Jeffreys wrote above looks like this,

$$P(A \,|\, B, \mathcal{M}_k) = \frac{P(AB \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)}$$

Supplemental Exercise 5.2.5: What is Jeffreys's culminating expression for the foundations of probability?

Solution to Supplemental Exercise 5.2.5:

Rather than immediately label his next development as *Bayes's Theorem*, he calls it rather the *principle of inverse probability*. I surmise in this time period the opprobrium attached to even mentioning the name of Bayes was so severe, that Jeffreys avoided the inevitable firestorm for as long as he possibly could.

In the following quote, I will intersperse my notational translation **in bold type** along the way, hoping this to be easier for the reader to follow than waiting until the end. Jeffreys's notation becomes even more a hindrance than before if you trying to quickly absorb his final result. I hope you will agree that my translation will really help you move through this quote at a much faster pace than if you had to stop and figure everything out on your own.

> **2.3.** The *principle of inverse probability* is an immediate consequence of the product rule. Let p be the initial data, θ a set of additional data, $q_1, q_2 \ldots q_n$ a set of hypotheses.
>
> **The "initial data p" here represent for Jeffreys all of the IP's encompassing background knowledge. This would be the rules of arithmetic, known physics, chemistry, biology, the language and connotations in which the problem is expressed, and so on to the exhaustion of what a typical IP might conceivably know.**
>
> **As mentioned, Jeffreys had previously used the symbol H for this global background knowledge, but not here. Jaynes used the symbols $\mathcal{I}$, or early on, the symbol X to represent such a concept.**
>
> **I choose not to use any symbol for this general background knowledge in my probability expressions because it seems to me that it would lead to an "infinite regression" of conditioning on all the elements in $\mathcal{I}$. I curtail such a regression right at the start by placing all of the burden of what is assumed to be known in the information inserted by $\mathcal{M}_k$.**

Jeffreys introduces a new symbol θ for "additional data," which really are any observational data. Therefore, $\theta \equiv \mathcal{D}$ in my notation.

The "set of hypotheses" are the same as the set of all models under consideration. Therefore, $q_1, q_2 \ldots q_n \equiv \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{\mathcal{M}}$. Now back to Jeffreys.

Then

$$P(q_r.\theta) \,|\, p) = P(q_r \,|\, p)\, P(\theta \,|\, q_r.p)$$

and also

$$= P(\theta \,|\, p)\, P(q_r \,|\, \theta.p)$$

These probability expressions become,

$$\boldsymbol{P(\mathcal{M}_k, \mathcal{D} \,|\, \mathcal{I}) = P(\mathcal{M}_k \,|\, \mathcal{I})\, P(\mathcal{D} \,|\, \mathcal{M}_k, \mathcal{I})}$$

and

$$\boldsymbol{P(\mathcal{M}_k, \mathcal{D} \,|\, \mathcal{I}) = P(\mathcal{D} \,|\, \mathcal{I})\, P(\mathcal{M}_k \,|\, \mathcal{D}, \mathcal{I})}$$

As Jeffreys says, he is applying the Product Rule. What he didn't say was that he was going to use the Commutativity axiom for the joint probability on the left hand side, and then apply the Product Rule twice.

Let's drop the conditioning on $\mathcal{I}$ at this point, and rearrange the terms to make everything much clearer,

$$\boldsymbol{P(\mathcal{D}, \mathcal{M}_k) = P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}$$

$$\boldsymbol{P(\mathcal{M}_k, \mathcal{D}) = P(\mathcal{M}_k \,|\, \mathcal{D})\, P(\mathcal{D})}$$

and therefore the ratio

$$\frac{P(q_r \,|\, \theta.p)}{P(q_r \,|\, p)\, P(\theta \,|\, q_r.p)} = \frac{1}{P(\theta \,|\, p)}$$

is the same for all q_r. This is the principle of inverse probability.

This unexpected, but obviously correct, expression translates to,

$$\boldsymbol{\frac{P(\mathcal{M}_k \,|\, \mathcal{D})}{P(\mathcal{M}_k)\, P(\mathcal{D} \,|\, \mathcal{M}_k)} = \frac{1}{P(\mathcal{D})}}$$

If the q_r are an exclusive and exhaustive set, the sum of the numerators of the expression on the left is 1, and

$$P(q_r \,|\, \theta.p) = \frac{P(q_r \,|\, p)\, P(\theta \,|\, q_r.p)}{\sum P(q_r \,|\, p)\, P(\theta \,|\, q_r.p)}$$

For Jeffreys, Bayes's Theorem is all about updating the prior probability of the models when conditioned on some data,

$$\boldsymbol{P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(\mathcal{D})}}$$

$$\boldsymbol{= \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}}$$

It is convenient to call $P(q_r \mid p)$ the prior probability of q_r, $P(q_r \mid \theta.p)$ the posterior probability, and $P(\theta \mid q_r.p)$ the likelihood. Then the theorem can be stated

$$\text{Posterior probability} \propto \text{prior probability} \times \text{likelihood}$$

The likelihood is the same as the 'direct probability' of many writers.

I deem it very important to emphasize exactly what Jeffreys called a "prior probability" and a "posterior probability." There can be no doubt whatsoever that the prior probability is the prior probability over model space, $P(\mathcal{M}_k)$. Likewise, there can be no doubt whatsoever that the posterior probability is the posterior probability over model space, $P(\mathcal{M}_k \mid \mathcal{D})$.

It is not that unusual for an agreement to be reached about the formal manipulation of symbols within some derivation as has happened here with Jeffreys and his prior probability $P(q_r \mid p)$ and his posterior probability $P(q_r \mid \theta.p)$ as compared to my $P(\mathcal{M}_k)$ and $P(\mathcal{M}_k \mid \mathcal{D})$.

Jeffreys might have enlightened us, and perhaps headed off the whole abortive enterprise of misinterpreting "prior and posterior probabilities," had he followed down the same path as Laplace in reasoning about "a prior probability over a set of hypotheses." But we know that didn't happen.

Instead, Jeffreys thought there was something seriously wrong when Bayes and Laplace assigned equal prior probability over a set of hypotheses at the very beginning of an inference before any data had accumulated. Interestingly, the end of Chapter 2 of his *Scientific Inference*, from which we have been quoting extensively, Jeffreys advances an alternative argument to Bayes and Laplace that ranked prior probabilities according to their *complexity*. Thus, simpler models would have a higher prior probability than more complex models. And one way to think about models along some complexity spectrum was how many parameters a model employed.

The reason I make such a big deal over this issue is because of the flagrant left turn by the Bayesian community, extending over very nearly the past hundred years, into the concept of "prior and posterior probabilities over parameters." Jeffreys was, in truth, a confirmed proponent for establishing probability distributions, both prior and posterior, over parameters.

I maintain that such a concept is ridiculous on the face of it (see Volumes II and III), and yet it is held to be a very respectable, albeit contentious, topic for Bayesians. It is heresy to suggest otherwise.

My resolution of this quandary, which is a core philosophical backbone supporting my presentation of inferencing, is that the "prior probability" represents a reasonable degree of belief in the particular information under some model. As a consequence, a probability assignment can be made for all the statements in the state space. It is therefore an IP's degree of belief in the truth of the assignment $Q_1, Q_2, \cdots, Q_n$ as made under the auspices of some model $\mathcal{M}_k$, made prior to, and independently of, the receipt of any observational data.

Supplemental Exercise 5.2.6: What did Jeffreys actually have to say about complexity and prior probability?

Solution to Supplemental Exercise 5.2.6:

To finish up this section, let's examine some final quotes from Jeffreys. These are very revealing because he amends what the reader might have assumed, given Jeffreys's development so far, as to what seemed to be a very clear definition of a "prior probability." He starts out with a disheartening disclosure about Bayes's Theorem involving an "yet unsolved and genuine difficulty." And you discover to your surprise that to Jeffreys a prior probability depends upon not only on a certain amount of information, but much previous observational information as well. So, he tells you that you must keep pushing further and further back to where you had no information at all. And now Jeffreys call the probabilities we must confront at this stage *initial probabilities.*

> **2.5** There is a genuine difficulty in making use of the principle of inverse probability, which is not yet completely solved, but there is no reason to believe it insoluble ... The point is this. At the beginning of any actual experiment we have a certain amount of information, which we have denoted by p. This usually includes much previous observational information, and it can be asked: what were the probabilities before you had that information? The scientist may be able to produce some less detailed information, but the same question can be repeated, and he can be driven to face the question: what were the probabilities before you had any observational evidence at all? (... I prefer to call them *initial probabilities.*)

We agree with Jeffreys that, at the start of any inference, we do condition the probability of any joint statement on the *information* in some model $\mathcal{M}_k$, as written $P(A = a_i \,|\, \mathcal{M}_k)$. Where we disagree is that the model $\mathcal{M}_k$ NEVER includes any previous observational data. So, we find ourselves by default right there at that initial stage of no observational data that Jeffreys labeled as "initial probabilities."

Apparently, however, Jeffreys's expression $P(r \,|\, p.H)$ may, and, in fact, does represent probabilities for r conditioned on any undefined amount of "information," observational or otherwise, in p and H.

Here is the crux of the matter. Jeffreys will not go down the path with Laplace on how to assign numerical values to "initial probabilities." Even though, there might be a "total lack of information due to any observational data," Jeffreys is predisposed to assign higher probabilities to simpler hypotheses, and consequently lower probabilities to more complex hypotheses.

> This amounts to saying that in the absence of observational evidence, the simpler law is the more probable and the initial probabilities can be placed in an order. To identify this with our postulate, all that we have to say is that the order of decreasing initial probabilities is that of increasing complexity.

In contrast, Laplace said to place equal probabilities on every single hypothesis, no matter how bizarre or compelling the hypothesis, no matter how simple or how complex the hypothesis.

I can imagine Laplace arguing that what else could the consequences of "total ignorance" or "complete lack of information" have on initial probabilities? Is the hypothesis of a two–headed or two–tailed coin simpler or more complex than the hypothesis of a fair coin? Do we assign a two–headed or two–tailed coin lower initial probabilities, or higher initial probabilities, than a fair coin?

If you know absolutely nothing about the physical make–up of the coin, why would any bias have any preference over any other? If you agree with Laplace, you never have to answer these difficult questions because all hypotheses are on an equal footing when the IP exists in the epistemological state of "complete ignorance."

The rules of probability theory, Jeffreys's own principle of inverse probability, provide any rearrangement of this equal standing over the set of hypotheses. The prior probabilities $P(\mathcal{M}_k)$, or Jeffreys's initial probabilities, get updated to posterior probabilities through "the principle of inverse probability" after data have been observed, through the simple expedient of following his own formal manipulation rules, namely, Bayes's Theorem in the form of,

$$P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(\mathcal{D})}$$

5.3 Solving Jaynes's Exercise 2.1

On page 34 of his book, Jaynes asks us to work out an exercise on our own.[1] Jaynes had just derived the formula for finding the inclusive OR of two statements, written in his notation as,

$$p(A + B \,|\, C) = p(A \,|\, C) + p(B \,|\, C) - p(AB \,|\, C) \qquad (2.66)$$

I had followed Jaynes's derivation fairly closely in section 5.4 of Volume I in a less–cluttered version written simply as,

$$P(A \vee B) = P(A) + P(B) - P(AB)$$

At this stage of my development of the formal manipulation rules, I had chosen to ignore any conditioning on a statement C, or, in my case, what should have been conditioning on a model $\mathcal{M}_k$.

[1] It is unclear whether Jaynes had actually solved in detail all of the exercises he posed in his book. Bretthorst made the comment in his **Editor's foreward** that some of these exercises functioned as placeholders for issues that Jaynes intended to address at some point later on, marked with the infamous "much more coming." But with his death, Bretthorst turned whatever Jaynes might have intended into these exercises.

Jaynes was asking whether the above very well–known formal manipulation rule could be turned around so that the IP was asking about the degree of belief in statement C as conditioned on A or B, or both, being true.

> **Exercise 2.1** Is it possible to find a general formula for $p(C \mid A+B)$, analogous to (2.66) from the product and sum rules? If so, derive it; if not explain why this cannot be done.

Supplemental Exercise 5.3.1: Answer Jaynes's question as posed above in his Exercise 2.1.

Solution to Supplemental Exercise 5.3.1:

I take the easy way out, albeit still proper I hope, of just using Bayes's Theorem right at the start instead of a longer rederivation involving the **Product Rule** and **Sum Rule**.

$$\begin{aligned}
P(C \mid A \vee B) &= \frac{P(CA \vee CB)}{P(A \vee B)} \\
&= \frac{P(AC \vee BC)}{P(A) + P(B) - P(AB)} \\
P(AC \vee BC) &= P(ABC) + P(A\overline{B}C) + P(ABC) + P(\overline{A}BC) - P(ACBC) \\
&= P(ABC) + P(A\overline{B}C) + P(\overline{A}BC) + P(ABC) - P(ABC) \\
&= P(ABC) + P(A\overline{B}C) + P(\overline{A}BC)
\end{aligned}$$

With the numerator of Bayes's Theorem taken care of, turn now to the denominator. It's always easier when expanding expressions like $P(A)$ and $P(B)$ to refer to a suitably constructed joint probability table. For a three statement joint probability table useful for this exercise, you could look ahead to Figure 6.1. Both $P(A)$ and $P(B)$, when expanded, will consist of four terms.

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

$$P(B) = P(ABC) + P(AB\overline{C}) + P(\overline{A}BC) + P(\overline{A}B\overline{C})$$

The sum $P(A) + P(B)$ will thus consist of eight terms. But two of these terms, namely $P(ABC)$ and $P(AB\overline{C})$, are repeated; therefore, they can be dropped. The formal manipulation rule will do this for us automatically when $P(AB) = P(ABC) + P(AB\overline{C})$ is subtracted.

The denominator in Bayes's Theorem will consist then of these six terms,

$$P(A \vee B) = P(ABC) + P(A\overline{B}C) + P(AB\overline{C}) + P(A\overline{B}\,\overline{C}) + P(\overline{A}BC) + P(\overline{A}B\overline{C})$$

The answer to Jaynes is then in the affirmative with a formula involving the following probabilities,

$$P(C \mid A \vee B) = \frac{P(ABC) + P(A\overline{B}C) + P(\overline{A}BC)}{P(ABC) + P(A\overline{B}C) + P(\overline{A}BC) + P(AB\overline{C}) + P(A\overline{B}\,\overline{C}) + P(\overline{A}B\overline{C})}$$

Again, it might be easier just to write down Bayes's Theorem with the cell locations from the eight cell joint probability table,

$$P(C \mid A \vee B) = \frac{\text{Cell 1} + \text{Cell 2} + \text{Cell 5}}{(\text{Cell 1} + \text{Cell 2} + \text{Cell 5}) + (\text{Cell 3} + \text{Cell 4} + \text{Cell 7})}$$

We conclude by remarking on two facts. First, given that conditioning on the truth of $A \vee B$ is the same as conditioning on the information in a particular model, cell 6 and cell 8 of the joint probability table must contain 0s. Under the information from such a model, A and B cannot both be false. (See Figure 6.1 in Supplemental Exercise 6.3.3 for the layout of the joint probability table.) Second, note the persistence of a pattern in the above version of Bayes's Theorem as we might expect from examination of the simplest generic version of Bayes's Theorem,

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(AB)}{P(AB) + P(\overline{A}B)}$$

That is, the negation of C must appear in the second set of all three terms of the denominator, cells 3, 4, and 7, after the first three terms that are the same as the numerator, cells 1, 2, and 5.

5.4 Solving Jaynes's Exercise 2.2

Immediately after posing Exercise 2.1, and without any intervening explanation, a follow–on Exercise 2.2 appears. It asks us to generalize the results from Exercise 2.1, so presumably, Jaynes thought that the initial pump priming exercise could indeed be answered in the affirmative.

Here is Exercise 2.2, and my response. It takes a slightly different interpretation than Jaynes, and when I am finished with my solution you will discern why.

> **Exercise 2.2.** Now suppose we have a set of propositions $\{A_1, \ldots, A_n\}$ which on information X are mutually exclusive: $p(A_iA_j \mid X) = p(A_i)\delta_{ij}$. Show that $p(C \mid (A_1 + A_2 + \cdots + A_nX)$ [sic] is a weighted average of the separate plausibilities $p(C \mid A_iX)$:

$$p(C \mid (A_1 + A_2 + \cdots + A_nX)[\text{sic}] = p(C \mid A_1X + A_2X + \cdots + A_nX)$$

$$= \frac{\sum_i p(A_i \mid X)\, p(C \mid A_iX)}{\sum_i p(A_i \mid X)} \qquad (2.67)$$

Supplemental Exercise 5.4.1: Solve Jaynes's Exercise 2.2, but perhaps not in the way he intended.

Solution to Supplemental Exercise 5.4.1:

Just as in the approach to Exercise 2.1, I will apply Bayes's Theorem right at the outset. But before I do that, and far more importantly, I impose this altered conceptual framework, together with its implied changed notation, on the problem.

I interpret the set of propositions $\{A_1, \ldots, A_n\}$ as the set of all models making numerical assignments to the probability for statement C. By definition, this set of models is mutually exclusive and, moreover, they are exhaustive.

Jaynes's statement C then becomes my generic statement A that the IP is uncertain about. My equivalent expression is then,

$$P(A \,|\, \mathcal{M}_1 \vee \mathcal{M}_2 \vee \cdots \mathcal{M}_k \vee \cdots \vee \mathcal{M}_{\mathcal{M}}) \equiv p(C \,|\, (A_1 + A_2 + \cdots + A_n X))$$

After this reorientation, Bayes's Theorem can be applied,

$$P(A \,|\, \mathcal{M}_1 \vee \cdots \vee \mathcal{M}_{\mathcal{M}}) = \frac{P(A\mathcal{M}_1 \vee \cdots \vee A\mathcal{M}_{\mathcal{M}})}{P(\mathcal{M}_1 \vee \cdots \vee \mathcal{M}_{\mathcal{M}})}$$

By the mutually exclusive nature of the models, the distribution of the probability symbol in the numerator and the denominator changes Bayes's Theorem to look like,

$$P(A \,|\, \mathcal{M}_1 \vee \cdots \vee \mathcal{M}_{\mathcal{M}}) = \frac{P(A\,\mathcal{M}_1) + P(A\,\mathcal{M}_2) + \cdots + P(A\,\mathcal{M}_{\mathcal{M}})}{P(\mathcal{M}_1) + P(\mathcal{M}_2) + \cdots + P(\mathcal{M}_{\mathcal{M}})}$$

Invoking the **Product Rule**, change the numerator to,

$$P(A \,|\, \mathcal{M}_1 \vee \cdots \vee \mathcal{M}_{\mathcal{M}}) = \frac{P(A \,|\, \mathcal{M}_1)\, P(\mathcal{M}_1) + P(A \,|\, \mathcal{M}_2)\, P(\mathcal{M}_2) + \cdots}{P(\mathcal{M}_1) + P(\mathcal{M}_2) + \cdots + P(\mathcal{M}_{\mathcal{M}})}$$

$$= \frac{\sum_{k=1}^{\mathcal{M}} P(A \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{M}_k)}$$

By the exhaustive property, the denominator in Bayes's Theorem is equal to 1. The final solution, as I would like to express it, becomes,

$$P(A) = \sum_{k=1}^{\mathcal{M}} P(A \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

In the end, I agree with Jaynes's verbal assessment that what we want is a *weighted average* of the individual degrees of belief about A. The interpretation here is that the probability of statement A conditioned on all of the background information X used in setting up the problem is a weighted average of the separate numerical assignments provided by the information from each model with respect to the weights thought of as the prior probabilities for the models.

Chapter 6

Bayes's Theorem

6.1 Status of Bayes's Theorem

I began the discussion of Bayes's Theorem in Volume I's Chapter Six with these three equivalent versions,

$$P(A \,|\, B) = \frac{P(AB)}{P(B)}$$

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

$$P(A \,|\, B) = \frac{P(B \,|\, A)\, P(A)}{P(B \,|\, A)\, P(A) + P(B \,|\, \overline{A})\, P(\overline{A})}$$

At the outset, I wanted versions of Bayes's Theorem with expressions in the simplest possible format. As a consequence, there were only two statements, with no indication of a model $\mathcal{M}_k$ assigning numerical values to the joint probabilities.

Furthermore, no symbol $\mathcal{D}$ appeared in any of these expressions, so there was no sense of any "data" involved in this preliminary introduction to Bayes's Theorem. Finally, there was no mention of any series of statements taking place over time, along the lines of finding, say, the probability of six HEADS and four TAILS in ten tosses of a coin.

My subsequent treatment of Bayes's Theorem in Chapter Six chose to focus on what happens when conditioning on any of the 16 logic functions. This was a prelude to a more expansive development in Chapter Seven examining the feasibility of using probability to generalize Classical Logic. In Chapter Ten, I persisted in this theme by showing how conditioning on logic functions and reliance on Bayes's Theorem allowed us to develop disciplined solutions to "brain twisters" whose underlying rationale on probability theory we could actually understand.

Wolfram's 256 elementary cellular automata were treated in depth in Chapters Eight and Nine first as deterministic machines, and secondly from the perspective of what happens when there is a loss of information on the IP's part about the ontology of the deterministic machines. It was seen that this kind of analysis could be carried out without deviating too much from the simple introductory approach adopted for Bayes's Theorem, other than increasing the number of statements and explicitly mentioning one model.

The concept here was that conditioning on the truth of some logic function, or equivalently in this case to some CA rule number, could be interpreted as being the same as conditioning on the truth of some model. Both would assign numerical values to abstract joint probabilities with logic functions possessing the interesting feature of inserting 0s at various places in the joint probability table.

In these supplemental exercises for Chapter Six, I will take the opportunity to add some more introductory forms of Bayes's Theorem to the original three shown above. After a few numerical exercises of these alternative forms, I would like to show further alternative versions of Bayes's Theorem that, in my opinion, possess a rather bizarre rationale. However, I wouldn't be including them if I didn't think they weren't instructive in some peculiar, but nonetheless interesting, way.

After that has been accomplished, I want to talk about something that bothers me a little bit from both a psychological and pedagogical perspective. It is Jaynes's very early treatment of probability calculations in his book. The first thing that bothers me is that I would have expected him to begin with an explication of Bayes's Theorem with numerical examples after the groundwork he had laid down in his first two Chapters. But he didn't do that. As a matter of fact, he never really gets into Bayes's Theorem at a foundational level for the entire rest of the book! This is curiously strange indeed.

As was evident in my preferred approach taken in Volume I, it was thought best to begin with elementary examples in Chapter Six, eschewing any complicating curlicues. After familiarity with the initial important concepts inherent in Bayes's Theorem, I started to explore less simple problems in Chapters Eleven through Fourteen, and in Chapter Sixteen. But even here, the probability that a statement was true at any given trial was defined to be not changing from trial to trial.

But Jaynes's Chapter 3 is entitled "**Elementary sampling theory**." Instead of delving into Bayes's Theorem, we are treated to a long, complicated discussion of drawing balls from an urn culminating in the hypergeometric distribution. In this case it seems plausible that the probabilities of drawing differently colored balls from a finite population in the urn are changing at each and every trial. Why begin numerical examples with the more difficult scenario of sampling without replacement when it seems to me it is easier to begin by talking about sampling with replacement? Therefore, why not the binomial distribution instead of the hypergeometric distribution for an introductory example?

Jaynes recognizes that what he labels as a "sampling distribution" is the same as what early Bayesians called the "direct probability." But by initially forcing our attention onto the "sampling distribution" all of the other constituents of Bayes's Theorem were, *de facto*, ignored.

For example, Harold Jeffreys used the language of "direct probabilities" in his initial treatment of Bayes's Theorem and remarked that the whole point of Bayes's Theorem was to calculate an "inverse probability."

Relying upon our third expression,

$$P(A \,|\, B) = \frac{P(B \,|\, A)\, P(A)}{P(B \,|\, A)\, P(A) + P(B \,|\, \overline{A})\, P(\overline{A})}$$

$P(A \,|\, B)$ on the left hand side is the inverse probability. The direct probability, Jaynes's "sampling probability," is the first term in the numerator, $P(B \,|\, A)$. During this whole process, the role of $P(A)$, as well as the denominator, are seemingly shunted aside.

Supplemental Exercise 6.1.1: Present some further instantiations of Bayes's Theorem in addition to the first three given in the opening.

Solution to Supplemental Exercise 6.1.1:

If statement A could be judged as TRUE for n_A different possible observations, then given that statement B has been observed in one of its possibilities as b_k, the probability that statement A will be observed as its jth possibility is calculated via Bayes's Theorem,

$$P(A = a_j \,|\, B = b_k) = \frac{P(B = b_k \,|A = a_j)\, P(A = a_j)}{\sum_{j=1}^{n_A} P(B = b_k \,|A = a_j)\, P(A = a_j)}$$

If the IP has constructed a 2×2 joint probability table to cover the case for all four joint statements of A and B, with definite numerical assignments Q_i in each cell of the table, then Bayes's Theorem for A TRUE given B FALSE might be presented as,

$$P(A \,|\, \overline{B}) = \frac{Q_3}{Q_3 + Q_4}$$

An alternative form useful for logistic regression (see Chapter Twenty Three, Volume II) with the statement to be predicted a binary variable as it is here,

$$P(A \,|\, \overline{B}) = \frac{1}{1 + \frac{Q_4}{Q_3}}$$

If the MEP is being used to make the numerical assignments to the Q_i (the exposition in Volume II), then the above expression turns into,

$$P(A \,|\, \overline{B}) = \frac{1}{1 + \frac{Q_4}{Q_3}} = \frac{1}{1 + \exp\{\, \sum_{j=1}^{m} \lambda_j \,[\, F_j(X = x_4) - F_j(X = x_3)\,]\,\}}$$

If the posterior predictive formula is used for an uninformed IP prior to any data, that is, $\sum_{i=1}^{4} N_i = N = 0$, then Bayes's Theorem produces,

$$P(A \,|\, \overline{B}, \mathcal{D}) = \frac{N_3 + 1}{(N_3 + 1) + (N_4 + 1)} = \frac{1}{2}$$

Please take care to note that this probability of 1/2 is conceptually distinct from an application of Bayes's Theorem conditioned on one specific model, say, where $Q_3 = Q_4 = 1/4$,

$$P(A \,|\, \overline{B}, \mathcal{M}_k) = \frac{1/4}{1/4 + 1/4} = \frac{1}{2}$$

If the IP wants to include a third statement $\mathcal{M}_k$ about some model assigning "correct" numerical values to all of the n joint probabilities in addition to the statements A and B, Bayes's Theorem can be written as,

$$\begin{aligned} P(A \,|\, B, \mathcal{M}_k) &= \frac{P(A, B, \mathcal{M}_k)}{P(B, \mathcal{M}_k)} \\ &= \frac{P(A \,|\, B, \mathcal{M}_k) \times P(B \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)} \\ &= \frac{P(A \,|\, B, \mathcal{M}_k) \times P(B \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \\ &= \frac{P(A, B \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \end{aligned}$$

Generalizing now to three statements A, B, and C, each of which can only be observed in two states, Bayes's Theorem for the probability of C TRUE given that B is TRUE and A is FALSE,

$$P(C \,|\, B\overline{A}, \mathcal{M}_k) = \frac{P(CB\overline{A} \,|\, \mathcal{M}_k)}{P(B\overline{A} \,|\, \mathcal{M}_k)} = \frac{P(\overline{A}BC \,|\, \mathcal{M}_k)}{P(\overline{A}BC \,|\, \mathcal{M}_k) + P(\overline{A}B\overline{C} \,|\, \mathcal{M}_k)}$$

Supplemental Exercise 6.1.2: Furnish an example of the instantiation for Bayes's Theorem where a logic function acts as a model.

Solution to Supplemental Exercise 6.1.2:

This was the prime focus of Chapter Six, especially in the exercises. From the previous supplemental exercise, we have this version of Bayes's Theorem where the IP is explicitly conditioning on some model $\mathcal{M}_k$.

$$P(A \,|\, B, \mathcal{M}_k) = \frac{P(A, B \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)}$$

The information resident in the model will assign definite numerical values Q_i for the $n = 4$ cell joint probability table constructed for A and B. The information in

a logic function as a model may result in some number of zeroes being assigned as numerical values to joint probabilities. This may enable Bayes's Theorem to produce probabilities of 1 or 0, in other words, certainty about a statement conditioned on another statement.

Examine the `NAND` logic function since it will be appearing shortly in upcoming supplemental exercises. Find the degree of belief that A is FALSE given that B is TRUE and the logic function serving as the model for numerical assignments is $A \uparrow B$,

$$P(\overline{A} \,|\, B, A \uparrow B) = \frac{P(\overline{A}, B \,|\, A \uparrow B)}{P(B \,|\, A \uparrow B)}$$

The DNF, or minterm canonical form, expansion of the `NAND` logic function, determines the coefficient of the building–block function AB to be $f(T,T) = F$, while the coefficients of the other three building–block functions are all T. I think you would agree that this is a nice instantiation of what a **NOT AND** definition should adhere to. Thus, a zero will be assigned to Q_1 and $1/3$ could be assigned to Q_2, Q_3, and Q_4. Bayes's Theorem then works out to,

$$P(\overline{A} \,|\, B, A \uparrow B) = \frac{Q_2}{Q_2 + Q_1} = \frac{1/3}{1/3 + 0} = 1$$

The IP's degree of belief is that it is certain that A must be FALSE.

6.2 Alternative Versions of Bayes's Theorem

Anthony Garrett [8] wrote an article with some very interesting insights on the role of the logic function `NAND`. In Chapter Two of these Supplemental Exercises, we looked at what Jaynes's had to say about the `NAND` logic function in terms of replacing `AND`, `OR`, and `NOT` as an adequate set of operations. Garrett's goal was actually to demonstrate that the **Sum Rule** and the **Product Rule** could be derived from `NAND`.

Supplemental Exercise 6.2.1: Give another version of Bayes's Theorem.

Solution to Supplemental Exercise 6.2.1:

Following the direction laid out by Garrett, it is possible to present yet another derivation of Bayes's Theorem as inspired by the `NAND` logic function. In order to provide some alternate variety to the traditional forward lay out of the steps, I will work backwards from Bayes's Theorem to the `NAND` logic function.

Start out with the simplest, most generic form of Bayes's Theorem,

$$P(A \,|\, B) = \frac{P(AB)}{P(B)}$$

Does this expression eventually lead to the `NAND` logic function? Of course, the motivation for the upcoming presentation of the steps in a backward fashion only makes sense if you had first worked them out in the forward direction.

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

$$-P(A \mid B)\, P(B) = -P(AB)$$

$$-P(A \mid B)\, P(B) + P(AB) = 0$$

$$1 - P(A \mid B)\, P(B) + P(AB) = 1$$

$$[\, 1 - P(AB)\,] + P(AB) = 1$$

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

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

Working backwards from Bayes's Theorem has resulted in an equality at the last step, obviously correct when thinking in terms of the sum over all four cells of the joint probability table.

After applying Boole's Expansion Theorem to any two variable logic function, we have the familiar expression,

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

where the coefficients of the `NAND` logic function are equal to,

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

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

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

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

Therefore, the `NAND` function (colloquially, where A and B are not both TRUE) leads to,

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

where $P(AB)$ must equal 0 because of the coefficient F appearing in the function expansion. This is exactly the last line in the above series of steps. But, of course, every logic function excepting the `FALSE` function must satisfy this aspect revealed by Bayes's Theorem.

Supplemental Exercise 6.2.2: Discuss and derive yet another version of Bayes's Theorem.

Solution to Supplemental Exercise 6.2.2:

For your delectation and delight, I am now going to show you in great detail the strangest "derivation" of Bayes's Theorem that I have ever come across. It appeared in an article by Garrett [7] accompanying the one cited in the previous exercise.

Although (unnecessarily?) convoluted, following his derivation is good practice in using the formal manipulation rules of probability. Garrett's stated goal is to use "the minterm method" in order to write out Bayes's Theorem as a function of three variables x, y, and z. Thus, the generic version of Bayes's Theorem could be expressed as,

$$P(A \,|\, B, \mathcal{M}_k) \equiv f(x, y, z)$$

In the derivation, Garret makes use of expansions like $A = AB \vee A\overline{B}$, and the distribution of probability over such expansions,

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

Garrett starts out his derivation by selecting the probabilities he is going to use as the arguments x, y, and z to his "Bayes's Theorem" function $f(x, y, z)$ as,

$$P(A \,|\, B, \mathcal{M}_k) = f\{P(A \,|\, \mathcal{M}_k), P(B \,|\, A, \mathcal{M}_k), P(B \,|\, \overline{A}, \mathcal{M}_k)\}$$

Garrett doesn't tell us the very critical rationale by which he initially picked these three arguments other than already knowing what the outcome was going to be.

Now, the interesting part of the derivation commences with Garrett's statement that,

> Before using the minterm method we must employ the product rule to ensure that every proposition is identically conditioned. ... In this problem the proposition I is the highest common (logical) factor to the right of the conditioning soliduses. Therefore we use the product rule to remove all other conditioning information to the left of the soliduses, with result ...

Garrett's proposition I is the same as my statement $\mathcal{M}_k$. What he is saying is that all statements involving A and B will be moved to the left of the conditioned upon symbol, while retaining only the model statement $\mathcal{M}_k$ to the right. So, for example, on the left hand side we must change $P(A \,|\, B, \mathcal{M}_k)$ to the equivalent,

$$\frac{P(AB \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)}$$

which does seem to be begging the question somewhat. But, as Garret says, he is only using the **Product Rule**.

Do the same thing to the three arguments of the function f on the right hand side with the result,

$$\frac{P(AB\,|\,\mathcal{M}_k)}{P(B\,|\,\mathcal{M}_k)} = f\left\{P(A\,|\,\mathcal{M}_k), \frac{P(AB\,|\,\mathcal{M}_k)}{P(A\,|\,\mathcal{M}_k)}, \frac{P(\overline{A}B\,|\,\mathcal{M}_k)}{P(\overline{A}\,|\,\mathcal{M}_k)}\right\}$$

The first argument x is already in the desired format so it need undergo no change. All three arguments are now identically conditioned solely on $\mathcal{M}_k$. All other conditioning statements have been moved to the left of the conditioned upon symbol.

At this juncture, Garrett tells us to make a "minterm expansion" for every statement we have moved to the left of the conditioned upon symbol. This is easily enough done, and we are left with,

$$\frac{P(AB\,|\,\mathcal{M}_k)}{P(AB \vee \overline{A}B\,|\,\mathcal{M}_k)} = f\left\{P(AB \vee A\overline{B}\,|\,\mathcal{M}_k), \frac{P(AB\,|\,\mathcal{M}_k)}{P(AB \vee A\overline{B}\,|\,\mathcal{M}_k)}, \frac{P(\overline{A}B\,|\,\mathcal{M}_k)}{P(\overline{A}B \vee \overline{A}\,\overline{B}|\,\mathcal{M}_k)}\right\}$$

The next step is easy as well. We just distribute the P operator over the $\vee$ symbol wherever it appears in any of the x, y, z arguments.

$$\frac{P(AB\,|\,\mathcal{M}_k)}{P(AB\,|\,\mathcal{M}_k) + P(\overline{A}B\,|\,\mathcal{M}_k)} = f\,\{x, y, z\}$$

$$x = P(AB\,|\,\mathcal{M}_k) + P(A\overline{B}\,|\,\mathcal{M}_k)$$

$$y = \frac{P(AB\,|\,\mathcal{M}_k)}{P(AB\,|\,\mathcal{M}_k) + P(A\overline{B}\,|\,\mathcal{M}_k)}$$

$$z = \frac{P(\overline{A}B\,|\,\mathcal{M}_k)}{P(\overline{A}B\,|\,\mathcal{M}_k) + P(\overline{A}\,\overline{B}\,|\,\mathcal{M}_k)}$$

What would you do next? I think I would have been stumped at this point, not really knowing how to proceed any further. But Garrett has a facility for solving functional equations. This will become evident if you read some of his other work on the derivation of the maximum entropy principle.[1]

[1] If your brain capacity does not exceed mine, you will find this tough going indeed.

To proceed, he re–labels the probability for each of the four minterms as,

$$\alpha = P(AB \,|\, \mathcal{M}_k)$$

$$\beta = P(\overline{A}B \,|\, \mathcal{M}_k)$$

$$\gamma = P(A\overline{B} \,|\, \mathcal{M}_k)$$

$$\delta = P(\overline{A}\,\overline{B} \,|\, \mathcal{M}_k)$$

I prefer to use language like "the coefficients for the orthogonal building block functions", the $f(T,T)$, $f(T,F)$, $f(F,T)$, and $f(F,F)$ for Garrett's α, β, γ, and δ.

In any case, the probabilities appearing on the left and right hand side of the equation can be expressed as,

$$\frac{\alpha}{\alpha+\beta} = f\left(\alpha+\gamma, \frac{\alpha}{\alpha+\gamma}, \frac{\beta}{\beta+\delta}\right)$$

The probabilities summed over all four coefficients must equal 1, and therefore,

$$\alpha+\beta+\gamma+\delta = 1$$

From this, the denominator in argument z is,

$$\beta+\delta = 1-(\alpha+\gamma)$$

so that we now have,

$$\frac{\alpha}{\alpha+\beta} = f\left(\alpha+\gamma, \frac{\alpha}{\alpha+\gamma}, \frac{\beta}{1-(\alpha+\gamma)}\right)$$

The end is nearly in sight by matching up with the original arguments in f,

$$\alpha+\gamma = x$$

$$\frac{\alpha}{\alpha+\gamma} = y$$

$$\frac{\beta}{1-(\alpha+\gamma)} = z$$

The final bit of algebraic manipulations bring us to this strange version of Bayes's Theorem. First, there is a very convenient cancellation of terms giving us,

$$\frac{\alpha}{\alpha+\gamma} \times (\alpha+\gamma) \Longrightarrow yx = \alpha$$

Then,

$$z = \frac{\beta}{1-(\alpha+\gamma)}$$

$$x = \alpha + \gamma$$

$$z = \frac{\beta}{1-x}$$

$$\beta = (1-x)z$$

In conclusion,

$$f(x,y,z) = \frac{\alpha}{\alpha+\beta} = \frac{xy}{xy+(1-x)z}$$

$$P(A\,|\,B,\mathcal{M}_k) = \frac{P(A\,|\,\mathcal{M}_k)\,P(B\,|\,A,\mathcal{M}_k)}{P(A\,|\,\mathcal{M}_k)\,P(B\,|\,A,\mathcal{M}_k) + [\,1-P(A\,|\,\mathcal{M}_k)\,]\,P(B\,|\,\overline{A},\mathcal{M}_k)}$$

This is my final bit of commentary. If you already knew Bayes's Theorem, then Garrett's derivation can be motivated through a far more direct approach using **Commutativity**, together with the **Sum Rule** and **Product Rule**. Our objective, after all, is to practice using probability's formal manipulation rules.

Following Garrett, we have just shown Bayes's Theorem as,

$$f(x,y,z) = \frac{xy}{xy+(1-x)z}$$

But right at the very beginning we were told that,

$$x = P(A\,|\,\mathcal{M}_k)$$

$$y = P(B\,|\,A,\mathcal{M}_k)$$

$$z = P(B\,|\,\overline{A},\mathcal{M}_k)$$

Earlier I complained about the lack of rationale for Garrett's pre–determined choice of these three arguments. I think what Garrett means to imply is not that you must find three correct arguments at the outset, but rather that you are free to try any set of probabilities as arguments x, y, $\cdots$

Then it is up to you to prove, or not, as the case may be, that a synthesis of one probability like $P(A\,|\,B,\mathcal{M}_k)$ is possible from any other probabilities you are curious about like $P(A\,|\,\mathcal{M}_k)$, $P(B\,|\,A,\mathcal{M}_k)$, and $P(B\,|\,\overline{A},\mathcal{M}_k)$.

So starting out with the truth of Bayes's Theorem and proceeding through a series of steps justified by the above mentioned **Commutativity**, together with the **Sum Rule** and **Product Rule**, we end up with $f(x,y,z)$,

$$
\begin{aligned}
P(A \,|\, B, \mathcal{M}_k) &= \frac{P(AB \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \\
&= \frac{P(BA \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \\
&= \frac{P(B \,|\, A, \mathcal{M}_k)\, P(A \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \\
&= \frac{P(B \,|\, A, \mathcal{M}_k)\, P(A \,|\, \mathcal{M}_k)}{P(BA \,|\, \mathcal{M}_k) + P(B\overline{A} \,|\, \mathcal{M}_k)} \\
&= \frac{P(B \,|\, A, \mathcal{M}_k)\, P(A \,|\, \mathcal{M}_k)}{P(B \,|\, A, \mathcal{M}_k)\, P(A \,|\, \mathcal{M}_k) + P(B \,|\, \overline{A}, \mathcal{M}_k)\, P(\overline{A} \,|\, \mathcal{M}_k)} \\
&= \frac{P(B \,|\, A, \mathcal{M}_k)\, P(A \,|\, \mathcal{M}_k)}{P(B \,|\, A, \mathcal{M}_k)\, P(A \,|\, \mathcal{M}_k) + P(B \,|\, \overline{A}, \mathcal{M}_k)\, [\, 1 - P(A \,|\, \mathcal{M}_k)\,]} \\
&= \frac{yx}{yx + z(1-x)} \\
&= \frac{xy}{xy + (1-x)z}
\end{aligned}
$$

6.3 Logic Functions as Models

Did anything in the arguments employed in the last section strike you as odd? Is there a whiff of circularity hovering about? Bayes's Theorem apparently also possesses the implication that all four cells of the joint probability table must sum to 1 as we demonstrated in Supplemental Exercise 6.2.1. There would be some explaining in order if it were otherwise. But does this fact lead somehow exclusively to the `NAND` logic function?

Every logic function, except for the `FALSE` logic function, that serves as a model for a probability assignment must in the end satisfy the requirement that,

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

To illustrate this fact, the following exercise is a variation on Exercise 6.6.15, Volume I, that asked about the probability that B was FALSE conditioned on the `NOR` logic function.

Supplemental Exercise 6.3.1: Use an argument stemming from the DNF of the `NOR` logic function to make a probability assignment to the four cells of a joint probability table.

Solution to Supplemental Exercise 6.3.1:

Ask *Mathematica* for the DNF of the `NOR` function with,

```
BooleanConvert[Nor[a, b]] // FullForm
```

which returns **`And[Not[a], Not[b]]`**.

We know that *Mathematica*'s answer comes from applying Boole's Expansion Theorem.

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

where the coefficients of the `NOR` logic function (Colloquially, neither A nor B nor both TRUE) must equal,

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

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

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

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

The `NOR` logic function is thus the model that is inserting information into the probability distribution over the statements A and B. Since the first three cells of the joint probability table must have a 0 assigned from the above DNF, that leaves 1 as the probability assignment for the fourth and final cell of the joint probability table indexing the joint statement, "A is FALSE and B is FALSE."

The relationship derived from Bayes's Theorem must certainly be true,

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

$$0 + 0 + 0 + 1 = 1$$

If the IP asks for the probability that A is TRUE given that the `NOR` logic function is the model making the probability assignments, then an application of Bayes's Theorem results in,

$$P(A \,|\, \mathcal{M}_k \equiv \texttt{NOR} \text{ logic function}) = \frac{P(A\,\overline{A}\,\overline{B})}{P(\overline{A}\,\overline{B})} = \frac{0}{1} = 0$$

Alternatively, it is easy to simply add the relevant cells of the joint probability table under this particular model to assert that,

$$P(A \,|\, \mathcal{M}_k) = P(B \,|\, \mathcal{M}_k) = 0 \text{ and } P(\overline{A} \,|\, \mathcal{M}_k) = P(\overline{B} \,|\, \mathcal{M}_k) = 1$$

Supplemental Exercise 6.3.2: Is the probability that B is FALSE affected by assuming the information inherent in the `NAND` logic function?

Solution to Supplemental Exercise 6.3.2:

The IP would like to use Bayes's Theorem to evaluate,

$$P(\overline{B}\,|\,\mathcal{M}_k \equiv \texttt{NAND}\text{ logic function}) = P(\overline{B}\,|\,A \uparrow B)$$

Consulting Table 2.6, Volume I, for the minterm canonical form, or for the fully expanded DNF, of $f_{12}(A,B)$ the `NAND` logic function, we find that,

$$MCF\,[\,f_{12}(A,B)\,] = A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B}$$

Therefore, a numerical assignment of 0 will go into cell 1 of the joint probability table for $P(AB\,|\,\mathcal{M}_k)$. Substituting the DNF of `NAND` into Bayes's Theorem,

$$\begin{aligned}
P(\overline{B}\,|\,A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B}) &= \frac{P(\overline{B} \wedge [\,A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B}\,]\,)}{P(A\overline{B} \vee \overline{A}B \vee \overline{A}\,\overline{B})} \\
&= \frac{P(\overline{B}A\overline{B} \vee \overline{B}\,\overline{A}B \vee \overline{B}\,\overline{A}\,\overline{B})}{1} \\
&= P(A\overline{B} \vee \overline{A}\,\overline{B}) \\
&= P(\overline{B})
\end{aligned}$$

We see that the probability that B is FALSE is unaffected by conditioning on the `NAND` logic function.

In a previous exercise, the numerical assignment to the joint probabilities making up B FALSE,

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

were both assigned a value of 1/3. However, under the information from the `NAND` logic function acting as a model, these two joint probabilities might equally well be assigned values of, say, 1/12 and 7/12.

Supplemental Exercise 6.3.3: Illustrate summing over relevant cells of a joint probability table as an alternative to explicitly calculating Bayes's Theorem.

Solution to Supplemental Exercise 6.3.3:

Suppose that the IP wants to find the probability for A TRUE or B TRUE, but not both TRUE at the same time. Suppose further that the state space is defined over three binary statements A, B, and C. The one model inserting information into the

distribution of probabilities over the state space is Rule 110. Thus, symbolically, the IP indicates this probability as,

$$P(A \oplus B \,|\, \mathcal{M}_k \equiv \text{Rule 110 logic function})$$

Rather than using Bayes's Theorem to form complicated Boolean expressions in both the numerator and denominator, use the DNF for Rule 110 to place 0s into the appropriate cells of the eight cell joint probability table. As one instantiation of a general Rule 110 model, assign equal probabilities to the remaining cells.

Volume I spent a lot of time examining the DNF for Rule 110. As a refresher on the results, and the care you have to take when using **`BooleanConvert[ ]`** to find the fully expanded DNF, consider,

```
f = BooleanFunction[110, 3]
```

followed by,

```
BooleanConvert[f[a, b, c]] // TraditionalForm
```

which returns,

$$(\neg a \wedge b) \vee (b \wedge \neg c) \vee (\neg b \wedge c) \text{ or } \overline{A}B \vee B\overline{C} \vee \overline{B}C$$

However, we don't want this condensed form for the DNF, but rather the fully expanded form. Expand each of the above three terms into a total of six terms,

$$\overline{A}BC \vee \overline{A}B\overline{C} \vee AB\overline{C} \vee \overline{A}B\overline{C} \vee A\overline{B}C \vee \overline{A}\,\overline{B}C$$

where we observe that one term $\overline{A}B\overline{C}$ is repeated and therefore can be dropped resulting in a total of five terms. Re–order these terms so that they would appear in cells 2, 3, 5, 6, and 7 of the joint probability table.

$$A\overline{B}C \vee AB\overline{C} \vee \overline{A}BC \vee \overline{A}\,\overline{B}C \vee \overline{A}B\overline{C}$$

Now we know to place the three 0s in cells 1, 4, and 8. The five remaining cells split the probability of 1 evenly so that cells 2, 3, 5, 6, and 7 are assigned a probability of 1/5.

We now ask what cells constitute the `XOR`, the "exclusive OR" logic function? The criterion is where A is TRUE and B is FALSE over both levels of C, or where A is FALSE and B is TRUE over both levels of C. These would be the four cells 2, 4, 5, and 7. Sum the probabilities in these cells,

$$\begin{aligned} P(A\overline{B}C) + P(A\overline{B}\,\overline{C}) + P(\overline{A}BC) + P(\overline{A}B\overline{C}) &= P(A\overline{B}) + P(\overline{A}B) \\ &= 1/5 + 0 + 1/5 + 1/5 \\ &= 3/5 \end{aligned}$$

to find the probability,

$$P(A \oplus B \,|\, \mathcal{M}_k \equiv \text{Rule 110 logic function}) = 3/5$$

This is made clear by perusing the joint probability table appearing in Figure 6.1 below. Remember that *Mathematica* imposes a different cell ordering than my construction of the joint probability table through **`Tuples[{True, False}, 3]`**. As a consequence, C would be the column variable and B the row variable, instead of the reverse as seen in Figure 6.1.

	A			$\bar{A}$			
	B	$\bar{B}$	P(BC) Cells 1+5	B	$\bar{B}$		
C	P(ABC) **0** Cell 1	P($A\bar{B}C$) **1/5** Cell 2 ∗	P(AC)	C	P($\bar{A}BC$) **1/5** Cell 5∗	P($\bar{A}\bar{B}C$) **1/5** Cell 6	P($\bar{A}C$) P(C)
$\bar{C}$	P($AB\bar{C}$) **1/5** Cell 3	P($A\bar{B}\bar{C}$) **0** Cell 4 ∗	P($A\bar{C}$)	$\bar{C}$	P($\bar{A}B\bar{C}$) **1/5** Cell 7∗	P($\bar{A}\bar{B}\bar{C}$) **0** Cell 8	P($\bar{A}\bar{C}$) P($\bar{C}$)
	P(AB)	P($A\bar{B}$)	P(A)		P($\bar{A}B$) P(B)	P($\bar{A}\bar{B}$) P($\bar{B}$)	P($\bar{A}$) 1

Figure 6.1: *A joint probability table with probability assignments dictated by the information in Rule 110. The four cells marked with $\star$ are involved in finding the probability.*

Supplemental Exercise 6.3.4: Generate another example just like the one in the previous exercise that connects with Wolfram's shortest axiom for basic logic.

Solution to Supplemental Exercise 6.3.4:

This exercise has a curious status for me. One day I happened to be reading Stephen Wolfram's blog site wherein he was celebrating the 200th year anniversary of George Boole's birth. At one point, he mentioned that in 1999 he had discovered what was most likely the simplest possible axiom system for logic consisting of just a single axiom.

I had remembered reading about this in Wolfram's book, but as I have mentioned above, the drive to find the simplest possible set of axioms for logic just doesn't resonate with me. I happen to prefer the larger, more profligate set of axioms as discussed in Volume I's Chapter Two. It's easier for me to perceive and select the relevant axiom from the larger set in derivations using Boolean operations.

Nonetheless, many people, as evidenced by my citing Jaynes and Garrett as the two examples I am most familiar with, have shown just how powerful the `NAND` or `NOR` logic functions can be. Jaynes showed us the three expressions involving the `NAND` operator that could replace the `AND`, `OR`, and `NOT` functions.

Wolfram's simplest axiom is also an expression that solely involves the `NAND` operator written in this notation as $\overline{\wedge}$. (As he says, he could just as easily have chosen the `NOR` operator.)

The simplest axiom for basic logic looks like this [18, pg. 773],

$$((a \,\overline{\wedge}\, b) \,\overline{\wedge}\, c) \,\overline{\wedge}\, (a \,\overline{\wedge}\, ((a \,\overline{\wedge}\, c) \,\overline{\wedge}\, a)) = c$$

Do you understand how I might prefer a less opaque, more expanded set of axioms?

In any case, this axiom *qua* symbolic logic expression didn't seem to be that much different than any of the other expressions we have been dealing with. And, in any case, I wanted to see how it translated over into a joint probability table just as in the previous exercise.

One aspect of working on this exercise is to simply practice the art of *nesting* functions in *Mathematica*. So, paying attention to this, and following Wolfram's expression as the guide, write out,

```
Nand[Nand[Nand[a, b], c], Nand[Nand[a, Nand[a, c], a]]
```

to confirm that *Mathematica* evaluates to Wolfram's expression, parentheses and all.

After some time had passed, I was rechecking this statement when I realized that the *Mathematica* expression I had originally set up above DID NOT match Wolfram's parentheses! So I myself was not paying sufficient attention to nesting and ordering. The correct expression matching Wolfram's basic axiom should be,

```
Nand[Nand[Nand[a, b], c], Nand[a, Nand[Nand[a, c], a]]]
```

which now does match Wolfram's expression, parentheses and all.

The next step would be to check that a logical equivalency does, in fact, exist between the left and right hand sides of the axiom. To that objective, write,

```
TautologyQ[Equivalent[Nand[Nand[Nand[a, b], c],
                           Nand[a, Nand[Nand[a, c], a]]], c]]
```

not surprisingly returning **`True`**.

Next up is to find the DNF for Wolfram's concatenation of `NAND` functions. *Mathematica* evaluates, again to no one's surprise,

```
BooleanConvert[Nand[Nand[Nand[a, b], c],
                          Nand[a, Nand[Nand[a, c], a]]]]
```

the concise formulation of **`c`**.

But I require the fully expanded DNF so I can see in which cells of the joint probability table the 0s are to be placed. So, as usual, I ask for the truth table through,

```
BooleanTable[Nand[Nand[Nand[a, b], c],
                          Nand[a, Nand[Nand[a, c], a]]]]
```

returning the list,

```
{True, False, True, False, True, False, True, False}.
```

This result enables me to construct the fully expanded DNF in four terms as,

$$ABC \vee A\overline{B}C \vee \overline{A}BC \vee \overline{A}\,\overline{B}C$$

It's not too difficult to see that this expression must reduce to C through extraction of the common factor,

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

followed by,

$$C \wedge T = C$$

We arrive now at my main objective which was to construct the joint probability table from the information inserted by this particular logic function, Wolfram's simplest axiom system. Since this state space consists of three binary statements, an eight cell joint probability table must be constructed. The numerical assignments of probability to each of these eight cells must derive from the information in some model $\mathcal{M}_k$. In this case, that model is Wolfram's logic function.

We will place 0s in four cells as dictated by the DNF, and four non–zero values summing to one in the remaining four cells. The 0s will be placed into cells 2, 4, 6, and 8. The non–zero assignments will be placed into cells 1, 3, 5, and 7.

Here is where we have to be careful in how the joint probability table was constructed. For this example, I built the joint probability table so that the cells matched up with the ordering that *Mathematica* demands through,

```
Tuples[{True, False}, 3]
```

For example, the third element of the returned list, itself a list consisting of three elements, would be **{True, False, True}** which must be matched with **True**. This corresponds to making the joint statement $A\overline{B}C$ the third cell in the joint probability table. Thus, contrary to the way I usually construct the joint probability table, C and $\overline{C}$ are the columns and B and $\overline{B}$ the rows.

All of this is made clear in Figure 6.2 at the top of the next page. As we shall discover in the next exercise, Wolfram's simplest axiom is Rule 170 from the set of 256 elementary cellular automata. An asterisk indicates a place holder for

	A				$\bar{A}$			
	C	$\bar{C}$	P(BC) Cells 1+5		C	$\bar{C}$		
B	P(ABC) * Cell 1	P(AB$\bar{C}$) 0 Cell 2	P(AB)	B	P($\bar{A}$BC) * Cell 5	P($\bar{A}$B$\bar{C}$) 0 Cell 6	P($\bar{A}$B)	P(B)
$\bar{B}$	P(A$\bar{B}$C) * Cell 3	P(A$\bar{B}\bar{C}$) 0 Cell 4	P(A$\bar{B}$)	$\bar{B}$	P($\bar{A}\bar{B}$C) * Cell 7	P($\bar{A}\bar{B}\bar{C}$) 0 Cell 8	P($\bar{A}\bar{B}$)	P($\bar{B}$)
	P(AC)	P(A$\bar{C}$)	P(A)		P($\bar{A}$C)	P($\bar{A}\bar{C}$)	P($\bar{A}$)	
					P(C)	P($\bar{C}$)		1

Figure 6.2: *A joint probability table with probability assignments dictated by the information in Rule 170.*

the assigned non–zero probabilities from whatever further information might be provided. Inserting the maximum amount of missing information after the four 0s have been assigned would dictate an assignment of 1/4 in the remaining four cells. The right hand side of the axiom is verified in that $P(C) = 1$.

My basic conundrum is this: What makes this joint probability table, or Rule 170, so special? What makes Wolfram's simplest axiom so special? This is a sincere question on my part. I am not criticizing the axiom; I feel that I must be missing some deeper import.

It is the most fundamental axiom of probability theory that the cells of the joint probability table must sum to 1. All models must adhere to this fundamental axiom whether they have been inspired by logic functions or not. And furthermore, it is a platitude that Bayes's Theorem implies that all cells of the joint probability table must sum to 1. The logic functions *are* special in the sense that they insert 0s into certain cells of the joint probability table, with the vitally important consequence that allows probability theory to reproduce logical deductions.

Supplemental Exercise 6.3.5: Which elementary cellular automata (ECA) is Wolfram's simplest axiom?

Solution to Supplemental Exercise 6.3.5:

After finding out that **`BooleanTable[ ]`** returned the list,

```
{True, False, True, False, True, False, True, False}
```

the following mapping, shown in Table 6.1 at the top of the next page, can be produced allowing us to translate over into the binary number for this particular elementary cellular automaton.

Table 6.1: *Translating the truth table for Wolfram's simplest axiom into the correct elementary cellular automata rule number.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
T	F	T	F	T	F	T	F
2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
1	0	1	0	1	0	1	0
128	0	32	0	8	0	2	0
Rule number 170							

Supplemental Exercise 6.3.6: How many ECA have four zeroes in their joint probability tables?

Solution to Supplemental Exercise 6.3.6:

There are exactly,

$$\binom{8}{4} = 70$$

ECA with four zeroes scattered around somewhere in four cells of the eight cell joint probability table. One of these is Rule 170 with zeroes in cells 2, 4, 6, and 8. We will now proceed to examine two more ECA from the total of 70 with four zeroes.

Supplemental Exercise 6.3.7: Select two more "special" ECA just like Rule 170.

Solution to Supplemental Exercise 6.3.7:

It would seem by symmetry that there ought to be two more joint probability tables "just like" the one shown previously in Figure 6.2 for Rule 170 and Wolfram's simplest axiom. The joint probability table for Rule 170 possessed the distinguishing characteristic feature that $P(C) = 1 - P(\overline{C}) = 1$. What then are the two ECA that implement $P(A) = 1 - P(\overline{A}) = 1$ and $P(B) = 1 - P(\overline{B}) = 1$?

It is pretty straightforward to work back towards these two models. The way the joint probabilities have been constructed, $P(A) = 1$ will result from the sum of the assigned probabilities in cells 1 through 4. $P(B) = 1$ will result from the sum of the assigned probabilities in cells 1, 2, 5, and 6.

If there are non–zero probabilities in cells 1 through 4, then that means that zeroes were assigned to cells 5 through 8. The DNF for the corresponding logic function must have coefficients of F for the orthogonal building block functions representing these four cells. These joint statements in cells 5 through 8 are, in

order,

$$\overline{A}BC, \overline{A}B\overline{C}, \overline{A}\,\overline{B}C, \overline{A}\,\overline{B}\,\overline{C}$$

The binary number matching up with the 1s and 0s must then be,

$$11110000 = 128 + 64 + 32 + 16 + 0 + 0 + 0 + 0 = \text{ Rule 240}$$

Likewise, if there are non–zero probabilities in cells 1, 2, 5, and 6, then that means that zeroes were assigned to cells 3, 4, 7, and 8. Once again, the DNF for the corresponding logic function must have coefficients of F for the orthogonal building block functions representing these four cells. These joint statements in cells 3, 4, 7, and 8 are, in order,

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

The binary number matching up with the 1s and 0s must then be,

$$11001100 = 128 + 64 + 0 + 0 + 8 + 4 + 0 + 0 = \text{ Rule 204}$$

In summary, the three models inserting information into a probability distribution such that $P(A) = 1$, $P(B) = 1$, and $P(C) = 1$ are the ECA labeled as Rules 240, 204, and 170, respectively.

Supplemental Exercise 6.3.8: Have the IP work out an example of Bayes's Theorem under the information of Rule 170.

Solution to Supplemental Exercise 6.3.8:

Suppose the IP has been instructed to avail itself of Bayes's Theorem by asking for the probability that A is TRUE when conditioned on knowing that B and C are also TRUE. The IP dutifully writes down the expression for Bayes's Theorem as,

$$P(A \,|\, B, C) = \frac{P(ABC)}{P(BC)}$$

but then is gently reminded of the further assumption that the information in a particular model $\mathcal{M}_k$, namely Rule 170, is guiding the actual numerical values to the probability assignments in the eight cell joint probability table for the three binary statements A, B, and C. Otherwise, how would the IP even begin to solve for the probability with Bayes's Theorem without any actual numbers to substitute?

Bayes's Theorem is then amended to reflect this fact,

$$\begin{aligned} P(A \,|\, B, C, \mathcal{M}_k \equiv \text{ Rule 170}) &= \frac{P(ABC \,|\, \mathcal{M}_k)}{P(BC \,|\, \mathcal{M}_k)} \\ &= \frac{P(ABC \,|\, \mathcal{M}_k)}{P(ABC \,|\, \mathcal{M}_k) + P(\overline{A}BC \,|\, \mathcal{M}_k)} \end{aligned}$$

Consulting the joint probability table in Figure 6.2, the IP notices that the assigned probabilities from cells 1 and 5 are called for,

$$P(A \,|\, B, C, \mathcal{M}_k) = \frac{\text{Cell 1}}{\text{Cell 1} + \text{Cell 5}}$$

These are both non–zero probabilities as dictated by the DNF for Rule 170, so the placeholder $\star$ values might be any legitimate probability such as 0.10 and 0.35.

For the purposes of this exercise, further assume a simpler informational scenario where the information entropy of the probability distribution has been maximized after taking account of the four mandatory zeroes. Thus, all four of the non–zero assignments are equal to 1/4. Bayes's Theorem now easily yields up a probability equal to 1/2 that A is TRUE conditioned on knowing that both B and C are TRUE,

$$P(A \,|\, B, C, \mathcal{M}_k) = \frac{1/4}{1/4 + 1/4} = 1/2$$

The above probability expressions were conditioned on assuming C to be TRUE. But under the model, C is always going to be TRUE. So, if C disappears from the right of the conditioned upon symbol as being irrelevant, do we recover the same probability?

$$\begin{aligned}
P(A \,|\, B, \mathcal{M}_k) &= \frac{P(AB \,|\, \mathcal{M}_k)}{P(B \,|\, \mathcal{M}_k)} \\
&= \frac{P(ABC \,|\, \mathcal{M}_k) + P(AB\overline{C} \,|\, \mathcal{M}_k)}{P(ABC \,|\, \mathcal{M}_k) + P(AB\overline{C} \,|\, \mathcal{M}_k) + P(\overline{A}BC \,|\, \mathcal{M}_k) + P(\overline{A}B\overline{C} \,|\, \mathcal{M}_k)} \\
&= \frac{\text{Cell 1} + \text{Cell 2}}{\text{Cell 1} + \text{Cell 2} + \text{Cell 5} + \text{Cell 6}} \\
&= \frac{\text{Cell 1}}{\text{Cell 1} + \text{Cell 5}}
\end{aligned}$$

Calculating the probability of B conditioned on the truth of A and C is very similar. Assuming equiprobable assignments to the non–zero cells, it also equals 1/2.

$$\begin{aligned}
P(B \,|\, A, \mathcal{M}_k) &= \frac{P(AB \,|\, \mathcal{M}_k)}{P(A \,|\, \mathcal{M}_k)} \\
&= \frac{P(ABC \,|\, \mathcal{M}_k) + P(AB\overline{C} \,|\, \mathcal{M}_k)}{P(ABC \,|\, \mathcal{M}_k) + P(AB\overline{C} \,|\, \mathcal{M}_k) + P(A\overline{B}C \,|\, \mathcal{M}_k) + P(A\overline{B}\,\overline{C} \,|\, \mathcal{M}_k)} \\
&= \frac{\text{Cell 1} + \text{Cell 2}}{\text{Cell 1} + \text{Cell 2} + \text{Cell 3} + \text{Cell 4}} \\
&= \frac{\text{Cell 1}}{\text{Cell 1} + \text{Cell 3}}
\end{aligned}$$

And, as we have already pointed out, $P(C \,|\, \mathcal{M}_k) = 1$ where $\mathcal{M}_k$ is the information under the Rule 170 ECA.

What if the IP had wanted to calculate the probability of A TRUE conditioned on B TRUE and C FALSE? Expressed in the usual symbols as $P(A \,|\, B, \overline{C}, \mathcal{M}_k)$ such a probability calculation is not allowed by the formal rules of probability manipulation. Bayes's Theorem will signal its displeasure by a value of 0 in the denominator.

The formal rules forbid conditioning on the truth of a contradictory statement. The information in model $\mathcal{M}_k \equiv$ Rule 170 imposes the fact that $P(C \,|\, \mathcal{M}_k) = 1$. The statement C cannot be both TRUE and FALSE at the same time.

Exercising the formal manipulation rules of probability theory will always satisfy whatever foundational axioms from logic happen to be imposed. This is the less than the awe–inspiring lesson I take away from Wolfram's simplest logic axiom. But maybe I'm missing something.

6.4 Exercise 6.6.18 of Volume I

These two exercises serve to extend the example of Bayes's Theorem first presented as Exercise 6.6.18 in Volume I.

Permutation of the three letters A, B, and C results in six orderings. Due to the **Commutativity** and **Associativity** axioms, all six orderings viewed as logic expressions are equivalent. For example, the ordering ABC is equivalent to the ordering CBA. Observe the **Commutativity** and **Associativity** axioms applied where required.

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

The above might be considered as a condensed version of proving a theorem by literally following Wolfram's prescription of taking a string and making a series of legal substitutions to produce a new string. The theorem is the final string.

This "theorem" was built on repeated applications of just two axioms from Boolean Algebra. I mention this because wouldn't it be nice if all theorems in mathematics followed such a general template? Is this were possible, rather than constantly invoking some special "mathematical insight" at almost every step to achieve the desired goal, our mathematical lives would be much happier.

Jaynes had expressed a similar frustration in connection with how Feller proved many of his theorems in probability theory. And I agree with Jaynes that in reading Feller you must be prepared at every turn to witness some new feat of magic.

In any case, the six possible ways of writing out the probability for the joint statement involving statements A, B, and C must all be equal. From the above example, $P(ABC) = P(CBA)$.

Supplemental Exercise 6.4.1: Repeat the argument of Exercise 6.6.18.

Solution to Supplemental Exercise 6.4.1:

List these six equal joint probabilities after exposure to the **Product Rule**.

$$P(ABC) = P(A \mid BC)\, P(B \mid C)\, P(C)$$

$$P(BAC) = P(B \mid AC)\, P(A \mid C)\, P(C)$$

$$P(ACB) = P(A \mid CB)\, P(C \mid B)\, P(B)$$

$$P(CAB) = P(C \mid AB)\, P(A \mid B)\, P(B)$$

$$P(BCA) = P(B \mid CA)\, P(C \mid A)\, P(A)$$

$$P(CBA) = P(C \mid BA)\, P(B \mid A)\, P(A)$$

The first two expressions were used in Exercise 6.6.18 in order to illustrate how Bayes's Theorem arrived at,

$$P(B \mid AC) = \frac{P(A \mid BC)\, P(B \mid C)}{P(A \mid C)}$$

Do the same thing to the last two expressions in the above list to arrive at Bayes's Theorem in the form of,

$$P(B \mid CA) = \frac{P(C \mid BA)\, P(B \mid A)}{P(C \mid A)}$$

But,

$$P(B \mid AC) = P(B \mid CA)$$

with the consequence that,

$$\begin{aligned} P(B \mid AC) &= \frac{P(ABC)}{P(AC)} \\ &= \frac{P(A \mid BC)\, P(B \mid C)\, P(C)}{P(A \mid C)\, P(C)} \\ &= \frac{P(A \mid BC)\, P(B \mid C)}{P(A \mid C)} \end{aligned}$$

$$\frac{P(A \mid BC)\, P(B \mid C)}{P(A \mid C)} = \frac{P(C \mid BA)\, P(B \mid A)}{P(C \mid A)}$$

This last equality between different Bayes's Theorem expressions is one that, and I think you might agree with this sentiment, is not immediately obvious if it were just presented to you out of the blue. Nevertheless, the straightforward conversion using Bayes's Theorem all over again on each constituent term, followed by cancellation, and finishing up with the **Commutativity** and **Associativity** axioms, shows that they are, in fact, equivalent.

$$\frac{P(A\,|\,BC)\,P(B\,|\,C)}{P(A\,|\,C)} = \frac{\frac{P(ABC)}{P(BC)} \times \frac{P(BC)}{P(C)}}{\frac{P(AC)}{P(C)}}$$

$$= \frac{P(ABC)}{P(AC)}$$

$$\frac{P(C\,|\,BA)\,P(B\,|\,A)}{P(C\,|\,A)} = \frac{\frac{P(CBA)}{P(BA)} \times \frac{P(BA)}{P(A)}}{\frac{P(CA)}{P(A)}}$$

$$= \frac{P(CBA)}{P(CA)}$$

$$= \frac{P(ABC)}{P(AC)}$$

How easy would it be to use Wolfram's simplest axiom to duplicate the effects of the **Commutativity** and **Associativity** axioms?

Supplemental Exercise 6.4.2: Apply the findings from the last exercise to the situation where there are three observations about whether some statement occurred.

Solution to Supplemental Exercise 6.4.2:

Attach the subscripts 1, 2, and 3 to statement A in order to indicate the trial number where A was observed TRUE or observed FALSE. For the A, B, C of the previous exercise, we substitute A_1, $\overline{A}_2$, and A_3 for A is true at the first observation, false on the second observation, and true on the third observation.

We saw how,

$$P(ABC) = P(CBA) = P(C\,|\,BA)\,P(B\,|\,A)\,P(A)$$

first by the **Commutativity** and the **Associativity** axioms, and then by the **Product Rule**.

Likewise, when the IP writes down the degree of belief in the truth of the joint statement that A was true on trial 1, false on trial 2, and true on trial 3, we have

the same relationships,

$$P(A_1, \overline{A}_2, A_3) = P(A_3, \overline{A}_2, A_1) = P(A_3 \mid \overline{A}_2, A_1)\, P(\overline{A}_2 \mid A_1)\, P(A_1)$$

as well as,

$$P(A_3, \overline{A}_2, A_1) = P(\overline{A}_2, A_3, A_1) = P(\overline{A}_2 \mid A_3, A_1)\, P(A_3 \mid A_1)\, P(A_1)$$

Suppose that the IP wants to leverage Bayes's Theorem to find the probability that A was true on the final trial given that it was false on the second trial and true on the first trial. Bayes's Theorem may appear in this form,

$$P(A_3 \mid \overline{A}_2, A_1) = \frac{P(\overline{A}_2 \mid A_3, A_1)\, P(A_3 \mid A_1)}{P(\overline{A}_2 \mid A_1)}$$

I submit to you that, while the above symbolic expression is undoubtedly correct, any attempt to translate the probabilities of the conditioned statements on the right hand side of Bayes's Theorem gives rise to some serious head scratching. What is the IP's degree of belief that A is false on trial 2 given that it was true on trial 3? What is the IP's degree of belief that A is true on trial 3 given that it was true on trial 1?

The easiest way out depends partially on something that should have been done right at the very outset. An additional statement indicating the information inserted by some model must always be included in the initial set of statements.

If we now include conditioning on $\mathcal{M}_k$ in every expression,

$$P(A_3 \mid \overline{A}_2, A_1, \mathcal{M}_k) = \frac{P(\overline{A}_2 \mid A_3, A_1, \mathcal{M}_k)\, P(A_3 \mid A_1, \mathcal{M}_k)}{P(\overline{A}_2 \mid A_1, \mathcal{M}_k)}$$

the IP is now permitted to assert independence of a probability on any trial, **under the model,** from the occurrences on any previous trials.

But there is something else of critical importance. It is how the state space was defined at the outset. If the state space was defined with dimension $n = 2$, no joint probability table involving A_1, A_2, and A_3 exists. The IP *can* define a state space of dimension of $n = 8$ with joint probabilities assigned to A_t, A_{t-1}, and A_{t-2} if it so wishes. If that were done, then the joint probabilities $P(A_3, A_2, A_1)$ appearing in the numerator of Bayes's Theorem could be substituted for directly.

However, what is usually done in this case for pragmatic reasons is to simplify Bayes's Theorem under some specified model and under the defined state space of $n = 2$, into the following simpler expression,

$$\begin{aligned} P(A_3 \mid \overline{A}_2, A_1, \mathcal{M}_k) &= \frac{P(\overline{A}_2 \mid \mathcal{M}_k)\, P(A_3 \mid \mathcal{M}_k)}{P(\overline{A}_2 \mid \mathcal{M}_k)} \\ &= P(A_3 \mid \mathcal{M}_k) \end{aligned}$$

The degree of belief that A will be true at the third observation given that it was false on the second observation and true on the first observation is determined solely by whatever assignment is made that A is true by the model $\mathcal{M}_k$. It doesn't depend at all on the observations made at the first and second trials.

The pedagogical example at this juncture always seems to revert back to a coin toss. The probability for HEADS at the third toss is, under say the fair model,

$$P(\text{HEADS}_3 \,|\, \text{TAILS}_2, \text{HEADS}_1, \mathcal{M}_{\text{Fair}}) = P(\text{HEADS} \,|\, \mathcal{M}_{\text{Fair}}) = 1/2$$

and the fact that TAILS was observed at the second toss, and HEADS on the first toss has no bearing on the matter at all.

Take care to note, however, that we have not dealt fully with the implications of the complete symbolic probability expression through the above ruse of asserting independence. By this I mean that the formal rules are NOT telling us that we MUST adopt independent and identically distributed probabilities, only that we are permitted to do so given the dimension of the state space and the one model that is assumed true.

If the world is not that simple, then probability theory tells us that we must then construct an eight cell joint probability table for A_1, A_2, and A_3. We have to abandon the simplicity of the two cell probability table for A.

Now correlations may exist between the outcomes occurring at the different trials. The probability that HEADS will appear on the third toss can depend on the conditional probability of a TAILS on the second toss when HEADS has appeared on both the first and third toss.

This approach quickly becomes very confusing and, moreover, taking this path in its full generality leads almost immediately to a combinatorial explosion. The motivation to simplify such complicated conditional probability expressions becomes paramount.

6.5 Jaynes and Sampling without Replacement

This last exercise permits us to segue seamlessly into a relevant discussion of the **Product Rule** and **Bayes's Theorem** as Jaynes used them to explain "sampling without replacement." As noted in my beginning remarks to this Chapter, Jaynes's treatment of this topic left me somewhat puzzled.

Chapter 3 of Jaynes's book is titled **Elementary sampling theory**. The first section is devoted to a discussion of the notion of *sampling without replacement*. On page 51, Jaynes informs us that,

> The first applications of the theory given in this chapter are, to be sure, rather simple and naive compared with the serious scientific inference that we hope to achieve later. Nevertheless, our reason for considering them in

> close detail is not mere pedagogical form. Failure to understand the logic of these simplest applications has been one of the major factors retarding the progress of scientific inference — and therefore of science itself — for many decades. Therefore we urge the reader, even one who is familiar with elementary sampling theory, to digest the contents of this chapter carefully before proceeding to more complicated problems.

Heeding Jaynes's warning that we should thoroughly understand elementary sampling before advancing to even more complicated inferential problems, let's take a look at sampling without replacement from an urn containing some number of red and white balls.

This exercise therefore should, echoing Jaynes, be a simple scientific inference. I will endeavor to demonstrate, contrary to what Jaynes has assured in this regard, there are, as they say, traps for the unwary. I want to highlight that even in this most "simple and naive inference problem" of sampling without replacement there are conceptual issues in probability theory that must be squarely dealt with before proceeding on to those "more complicated problems."

Supplemental Exercise 6.5.1: Repeat some of the same arguments that ended up in the last section as an introduction to what Jaynes labeled as elementary sampling and leading eventually to the hypergeometric distribution.

Solution to Supplemental Exercise 6.5.1:

One core concept was not addressed in these beginning formulations of Bayes's Theorem. It concerns an extremely important aspect of how the IP is allowed to manipulate probability expressions.

Certainly an introductory version of Bayes's Theorem like,

$$P(A \,|\, B, \mathcal{M}_k) = \frac{P(A, B \,|\, \mathcal{M}_k)}{P(A, B \,|\, \mathcal{M}_k) + P(\overline{A}, B \,|\, \mathcal{M}_k)}$$

was adequate to our needs in the situation where logic functions were used as models. Once the state space of, say, dimension $n = 4$ has been defined for joint statements about A and B, and numerical values have been assigned to the joint probability of all four cells under some logic function, everything proceeds smoothly.

In fact, one of the virtues of this form of Bayes's Theorem was an immediate inference and degree of belief about statement A upon confirmation of statement B. There was no sense that vast amounts of data were required to achieve a degree of belief about A. Refer back to the **Life on Mars** scenario from Chapter Seven, Volume I, for an example of an immediate inference requiring only confirmation of the presence of water, a statement B, not requiring any data. Additional comments appear in current sections 7.3, 7.4, and Chapter Thirty Nine of Volume III.

However, scientific inference does depend on processing large amounts of data via Bayes's Theorem. How do we handle repeated trials? Generically, if we have N pieces of data, A_N, A_{N-1}, $\cdots$, A_t, $\cdots$, A_1, how can we assign numerical values to the joint probability? The **Product Rule** tells us that we could decompose the joint probability in full generality as,

$$P(A_N, A_{N-1}, \cdots, A_t, \cdots, A_1) = P(A_N \,|\, A_{N-1}, \cdots, A_t, \cdots, A_1) \times$$
$$P(A_{N-1} \,|\, A_{N-2}, \cdots, A_t, \cdots, A_1) \times \cdots \times P(A_2 \,|\, A_1) \times P(A_1)$$

But as the amount of data grows larger and larger and N increases, how can an IP possibly assign numerical values contingent on an ever changing set of data that occurred on previous trials?

The usual way out of this mess is through an assertion that the probability for each term depends only on the model and not on any of the past data. Thus, the joint probability simplifies greatly under this asumption to,

$$\begin{aligned} P(A_N, A_{N-1}, \cdots, A_t, \cdots, A_1, \mathcal{M}_k) &= P(A_N \,|\, \mathcal{M}_k) \times \cdots \times P(A_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \\ &= q_1^{N_1}\, q_2^{N_2} \cdots q_n^{N_n}\ \mathrm{pdf}\,(q_1, q_2, \cdots, q_n)\, \mathrm{d}q_i \end{aligned}$$

where n is the already defined dimension of the state space for statement A. Thus, any assigned Q_i are placed into a n cell probability table.

Now we come to the crux of the issue. Sampling without replacement does not (cannot?) rely upon this ruse. It continues to maintain its reliance on the decomposition effected by the **Product Rule** and,

$$\begin{aligned} P(A_N, A_{N-1}, \cdots, A_t, \cdots, A_1) = \sum_{k=1}^{\mathcal{M}} & P(A_N \,|\, A_{N-1}, \cdots, A_t, \cdots, A_1, \mathcal{M}_k) \times \\ & P(A_{N-1} \,|\, A_{N-2}, \cdots, A_t, \cdots, A_1, \mathcal{M}_k) \times \cdots \times \\ & P(A_2 \,|\, A_1, \mathcal{M}_k) \times P(A_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \end{aligned}$$

This means that each individual probability expression on the right hand side must be figured out by some logical or physical analysis for a probability that is dependent on all of the past data that have occurred. The solution that Jaynes discusses for this imbroglio is the classical scenario of drawing differently colored balls from an urn.

Supplemental Exercise 6.5.2: Recapitulate and augment Jaynes's version of elementary sampling theory without replacement through an easy numerical example.

Solution to Supplemental Exercise 6.5.2:

Calculate the probability for drawing exactly three red balls and two white balls in a total of five draws from an urn that contains a total of ten balls, seven of which are red, and three are white. The ultimate solution for this probability calculation leads to the hypergeometric distribution. Appendix C contains details of the hypergeometric distribution when discussed from the *Mathematica* perspective.

I now list the notation Jaynes used for elementary sampling without replacement together with the specific numerical values employed in my example.

$$r = 3$$

$$w = 2$$

$$n = r + w = 5$$

$$N = 10$$

$$M = 7$$

$$N - M = 3$$

You can appreciate that I am trying to force Jaynes's pedagogical example of sampling without replacement to be even more transparent by explicitly specifying these values. Using Jaynes's notation to start, suppose that statement B, "the Bernoulli urn scenario" is realized with a total of $N = 10$ red and white balls in the urn. There are $M = 7$ red balls, and $N - M = 3$ white balls. We will draw $n = 5$ balls from the urn.

The probability of a red ball on the first draw written as $P(R_1 \,|\, B)$, while the probability for a white ball on the third draw is written as $P(W_3 \,|\, B)$. The statement B is a full description of the Bernoulli urn scenario specifying, among other details, feasible values for N, M, and n. Most important of all, this all–encompassing statement B also serves as our model $\mathcal{M}_k$, because as we shall see, it makes a definite numerical assignment to the probabilities of drawing a red or white ball at any trial.

Now, it is clear from the **Sum Rule** that no matter what numerical assignments are made to the probabilities of drawing a red or white ball, these probabilities must sum to 1. And whether we write the generic $P(A \,|\, \mathcal{M}_k) + P(\overline{A} \,|\, \mathcal{M}_k) = 1$, or $P(R_j \,|\, B) + P(W_j \,|\, B) = 1$ for this scenario, the IP recognizes that, for the purposes of the **Sum Rule**, what information the probabilities of a red or white ball are conditioned on at any trial doesn't make any difference.

Here, that information is from Jaynes's description of statement B, the "Bernoulli urn scenario." In order to proceed further with the problem, Jaynes asserts that conditioning on B must also mean that,

$$P(R_1 \mid B) = \frac{M}{N} = \frac{7}{10} \text{ and } P(W_1 \mid B) = 1 - \frac{M}{N} = \frac{3}{10}$$

Furthermore, after that first red ball, the probabilities for the second ball become,

$$P(R_2 \mid R_1, B) = \frac{M-1}{N-1} = \frac{6}{9} \text{ and } P(W_2 \mid R_1, B) = \frac{N-M}{N-1} = \frac{3}{9}$$

I expected Jaynes to explicitly invoke Bayes's Theorem at this point. But he did not. Not having done so, these numerical assignments might be interpreted as merely the plausible ones based directly on the conditioning statements.

If Bayes's Theorem had been used,

$$P(R_2 \mid R_1, B) = \frac{P(R_2\, R_1 \mid B)}{P(R_1 \mid B)} = \frac{\frac{M-1}{N-1} \times \frac{M}{N}}{\frac{M}{N}} = \frac{M-1}{N-1}$$

It's instructive to stop here and critically analyze what Jaynes has just done.

Supplemental Exercise 6.5.3: Does Jaynes's statement B really allow him to make those assignments?

Solution to Supplemental Exercise 6.5.3:

There is nothing in statement B, and I mean statement B exactly as Jaynes wrote it down, categorically dictating an assignment of 7/10 to drawing a red ball from the urn. It is merely intuitively plausible that one very good assignment might be 7/10 based on the *physical* reality of 7 red balls and 3 white balls sitting in the urn. Furthermore, it is common sense to make a probability assignment that depends on decrementing the relevant number of balls by one after it is known what was drawn.

But unless the statement R_1 is explicitly conditioned on the information in a model $\mathcal{M}_k$ that asserts categorically $\mathcal{M}_k \rightarrow (q = 7/10, 1 - q = 3/10)$, R_1 cannot be assigned the numerical value of 7/10.

The frustrating aspect of all this is that Jaynes makes my point in a powerful and lucid manner when he says immediately after the assignment for the probability of a red ball on the first draw [11, pg. 52],

> Let us understand clearly what this means. The probability assignments ... are not assertions of any physical property of the urn or its contents; they are a description of the *state of knowledge* of the robot prior to the drawing. Indeed, were the robot's state of knowledge different from B as just defined (for example, if it knew the actual positions of the red and white balls in the urn, or if it did not know the true values of N and M), then its probability assignments for R_1 and W_1 would be different; but the real properties of the urn would be just the same.

Just so. But what is the assignment of $P(R_1 \,|\, B) = 7/10$ if not an assertion based on the physical property of this urn that it contains exactly 7 red balls and 3 white balls? Is not $P(R_1 \,|\, \mathcal{M}_k) = 0.69$ or 0.71 also a legitimate assignment that could be a description of the IP's state of knowledge?

Supplemental Exercise 6.5.4: Continue on with Jaynes's solution.

Solution to Supplemental Exercise 6.5.4:

Nevertheless, let's adopt the probability assignments from Jaynes's one model B. Then, the *joint* probability of a red ball on the second draw *and* a red ball on the first draw is calculated from the **Product Rule** as,

$$P(R_2\, R_1 \,|\, B) = P(R_2 \,|\, R_1, B)\, P(R_1 \,|\, B) = \frac{6}{9} \times \frac{7}{10}$$

where, once again, the *physical reality* of one less red ball and one less total number of balls in the urn is driving the assignment to $P(R_2 \,|R_1, B)$. However, notice the strong constraint that **Commutativity** and **Associativity** impose on probability expressions,

$$P(R_1\, R_2 \,|\, B) = P(R_1 \,|R_2, B)\, P(R_2 \,|\, B) = P(R_2\, R_1 \,|\, B) = \frac{6}{9} \times \frac{7}{10}$$

Are you starting to get a little confused with the possible implication that,

$$P(R_1 \,|R_2, B) = \frac{6}{9} \text{ and } P(R_2 \,|\, B) = \frac{7}{10}?$$

And so it continues in like manner for any situation. The joint probability for red ball on the third draw, white ball on the second draw, and a red ball on the first draw becomes,

$$P(R_3\, W_2\, R_1 \,|\, B) = P(R_3 \,|\, W_2, R_1, B)\, P(W_2 \,|\, R_1, B)\, P(R_1 \,|\, B) = \frac{6}{8} \times \frac{3}{9} \times \frac{7}{10}$$

Once again, this joint probability broken down by the **Product Rule** must be the same as when writing the joint probability in the equally acceptable order of,

$$P(R_1\, W_2\, R_3 \,|\, B) = P(R_1 \,|\, W_2, R_3, B)\, P(W_2 \,|\, R_3, B)\, P(R_3 \,|\, B) = \frac{6}{8} \times \frac{3}{9} \times \frac{7}{10}$$

From these simple numerical experiments, we can quickly discern, because of the **Commutativity** property of ordinary numbers under multiplication, that the factors making up the probability calculation could be rearranged to look like,

$$P(R_3\, W_2\, R_1 \,|\, B) = \frac{7 \times 6 \times 3}{10 \times 9 \times 8}$$

Supplemental Exercise 6.5.5: Continue on with this numerical example illustrating the derivation of Jaynes's general result that culminates in the hypergeometric distribution.

Solution to Supplemental Exercise 6.5.5:

Extrapolating to, say, the probability for a red ball on the first three draws, we now have,

$$P(R_1\, R_2\, R_3 \,|\, B) = \frac{7 \times 6 \times 5}{10 \times 9 \times 8}$$

Guided by this numerical example, write out a symbolic expression,

$$P(R_1\, R_2\, R_3 \,|\, B) = \frac{M \times (M-1) \times (M-2)}{N \times (N-1) \times (N-2)} = \frac{M!}{(M-3)!} \times \frac{(N-3)!}{N!}$$

Again, following Jaynes's notation, r is the number of red balls drawn from the urn, so write,

$$\begin{aligned} P(R_1\, R_2\, R_3 \,|\, B) &= \frac{M!}{(M-r)!} \times \frac{(N-r)!}{N!} \\ &= \frac{7!}{(7-3)!} \times \frac{(10-3)!}{10!} \\ &= \frac{7 \times 6 \times 5}{10 \times 9 \times 8} \\ &= \frac{7}{24} \end{aligned}$$

Involving the white balls is necessary at some point. Suppose that the fourth and fifth draws are white balls. We would then need to calculate $P(R_1\, R_2\, R_3\, W_4, W_5 \,|\, B)$. Following the same argument that Jaynes has been using all along, it would seem that this probability should be,

$$P(R_1\, R_2\, R_3\, W_4, W_5 \,|\, B) = \frac{7 \times 6 \times 5 \times 3 \times 2}{10 \times 9 \times 8 \times 7 \times 6}$$

First, let's gain some confidence from the numerical calculation using Jaynes's Equation (3.15) where we have the new symbols $r + w = n$ for the number of r red balls drawn, the number of w white balls drawn, and n, the total number of balls drawn, so that $r = 3$, $w = 2$, and $n = 5$,

$$
\begin{aligned}
P(R_1 \cdots R_r\, W_{r+1}, \cdots, W_n \,|\, B) &= \frac{M!\,(N-M)!\,(N-n)!}{(M-r)!\,(N-M-w)!\,N!} \\
&= \frac{7!\,(10-7)!\,(10-5)!}{(7-3)!\,(10-7-2)!\,10!} \\
&= \frac{7!\,3!\,5!}{4!\,10!} \\
&= \frac{7\times 6\times 5\times 3\times 2}{10\times 9\times 8\times 7\times 6}
\end{aligned}
$$

which is reassuring.

Let's now deconstruct how Jaynes arrived at this formula. Returning to the same arguments as seen in the previous exercise, what would be the probability of drawing out two white balls on the first two draws? Specifically under the numerical example,

$$P(W_1, W_2 \,|\, B) = \frac{3}{10} \times \frac{2}{9}$$

and more generally,

$$P(W_1, W_2 \,|\, B) = \frac{N-M}{N} \times \frac{N-M-1}{N-1}$$

and then even more generally for the probability of drawing w white balls,

$$
\begin{aligned}
P(W_1, W_2, \cdots, W_w \,|\, B) &= \frac{N-M}{N} \times \frac{N-M-1}{N-1} \times \cdots \times \frac{N-M-w+1}{N-w+1} \\
&= \frac{(N-M)!}{(N-M-w)!} \times \frac{(N-w)!}{N!}
\end{aligned}
$$

You must pay close attention to Jaynes at this juncture because he wants us to consider the probability that the white balls have been drawn after the red balls. More specifically, condition on the fact that r red balls have already been drawn, and the IP wants the probability for the $(r+1)^{th}$ through the $(r+w)^{th}$ white balls.

He writes this as $P(W_{r+1} \cdots W_{r+w} \,|\, R_1 \cdots R_r, B)$ because he then wants to use the **Product Rule** to find the joint probability for all $r + w = n$ draws, as in,

$$P(R_1 \cdots R_r\, W_{r+1} \cdots W_n \,|\, B) = P(W_{r+1}, \cdots, W_{r+w} \,|\, R_1, \cdots, R_r, B) \times P(R_1, \cdots, R_r \,|\, B)$$

We have already found the second term in the previous exercise, and almost the first term just above. That second term is,

$$P(R_1 \cdots R_r \,|\, B) = \frac{M!\,(N-r)!}{(M-r)!\,N!}$$

Jaynes reasons that the first conditional probability term must change to,

$$\frac{(N-M)!}{(N-M-w)!} \times \frac{(N-w)!}{N!} \longrightarrow \frac{(N-M)!}{(N-M-w)!} \times \frac{(N-w-r)!}{(N-r)!}$$

because conditioning on the truth of r red balls having been drawn, we must reduce the factorial terms by r.

Now it becomes a matter of multiplying these two factorial expressions and canceling the $(N-r)!$ terms,

$$\begin{aligned} P(R_1 \cdots R_r\, W_{r+1} \cdots W_n \,|\, B) &= \frac{(N-M)!}{(N-M-w)!} \times \frac{(N-w-r)!}{(N-r)!} \times \frac{M!\,(N-r)!}{(M-r)!\,N!} \\ &= \frac{(N-M)!\,(N-n)!\,M!}{(N-M-w)!\,(M-r)!\,N!} \end{aligned}$$

The $(N-n)$ term arises because of,

$$w + r = n \longrightarrow -w - r = -n$$

Supplemental Exercise 6.5.6: What is the final step that results in the hypergeometric distribution?

Solution to Supplemental Exercise 6.5.6:

At the end of the last exercise, we had managed to derive Jaynes's formula for any *one* particular sequence of r red and w white balls. The numerical verification used the particular sequence of $r = 3$ red balls on the first three draws and $w = 2$ white balls on the fourth and fifth draws for the total of $n = 5$ draws from the urn. Together with the stipulation of $N = 10$ balls in the urn, $M = 7$ red and $(N-M) = 3$ white, the numbers coming out of the factorial expressions were seen to be,

$$\begin{aligned} P(R_1\, R_2\, R_3\, W_4\, W_5 \,|\, B) &= \frac{(N-M)!\,(N-n)!\,M!}{(N-M-w)!\,(M-r)!\,N!} \\ &= \frac{3!\,5!\,7!}{1!\,4!\,10!} \\ &= \frac{7!}{4!} \times 3! \times \frac{5!}{10!} \\ &= \frac{7 \times 6 \times 5 \times 3 \times 2}{10 \times 9 \times 8 \times 7 \times 6} \end{aligned}$$

This is the same as Jaynes's Equation (3.17),

$$P(R_1 \cdots R_r\, W_{r+1} \cdots W_n \,|\, B) = \frac{M(M-1)\cdots(M-r+1)(N-M)(N-M-1)\cdots(N-M-w+1)}{N(N-1)\cdots(N-n+1)}$$

$$= \frac{7 \times 6 \times 5 \times 3 \times 2}{10 \times 9 \times 8 \times 7 \times 6}$$

where he emphasized that these same numbers in the numerator and denominator are going to show up for the probability of any particular sequence of three red and two white balls.

Therefore, the final step is to multiply the probability for any particular sequence by the binomial coefficient,

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} = \binom{5}{3} = 10$$

The probability for three red balls and two balls in any order in five draws from the urn is then,

$$P(3 \text{ red, } 2 \text{ white} \,|\, B) = 10 \times \frac{7 \times 6 \times 5 \times 3 \times 2}{10 \times 9 \times 8 \times 7 \times 6} = \frac{5}{12}$$

The general result for the formula of the hypergeometric distribution is then, as Jaynes shows in his Equation (3.22),

$$P(A \,|\, B) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}$$

where A is the statement "Exactly r red balls in n draws in any order."

The task before us then is to see whether all of the factorials in the multiplication of the probability of a particular sequence by the total number of such sequences can be rearranged into the neat package of two binomial coefficients in the numerator over the one binomial coefficient in the denominator as Jaynes specified in his Equation (3.22),

$$P(A \,|\, B) = \frac{n!}{r!\,(n-r)!} \times \frac{(N-M)!\,(N-n)!\,M!}{(N-M-w)!\,(M-r)!\,N!}$$

The most obvious place to start would be,

$$\binom{M}{r} = \frac{M!}{r!\,(M-r)!}$$

to yield currently,

$$P(A \,|\, B) = \binom{M}{r} \times \frac{n!}{(n-r)!} \times \frac{(N-M)!\,(N-n)!}{(N-M-w)!\,N!}$$

Rearrange this expression to place next to each other the set of factorial terms for the second binomial coefficient, and where we know that $w = n - r$,

$$\binom{M}{r} \times \frac{n!}{(n-r)!} \times \frac{(N-M)!\,(N-n)!}{(N-M-w)!\,N!} = \binom{M}{r} \times n! \times \frac{(N-M)!}{(N-M-(n-r))!\,(n-r)!} \times \frac{(N-n)!}{N!}$$

where we can replace the third term on the right hand side with what we want,

$$P(A\,|\,B) = \binom{M}{r} \times n! \times \binom{N-M}{n-r} \times \frac{(N-n)!}{N!}$$

Rearranging once again,

$$P(A\,|\,B) = \binom{M}{r} \times \binom{N-M}{n-r} \times \frac{(N-n)!\,n!}{N!}$$

Since the third term on the right hand side is obviously,

$$\frac{(N-n)!\,n!}{N!} = \frac{1}{\binom{N}{n}}$$

we have managed to reproduce Jaynes's Equation (3.22) for the hypergeometric distribution,

$$P(A\,|\,B) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}$$

Supplemental Exercise 6.5.7: Have *Mathematica* numerically verify these efforts in the previous supplemental exercises.

Solution to Supplemental Exercise 6.5.7:

Having followed Jaynes's derivation for a sampling scenario of drawing balls from an urn without replacement, we are able to confirm that it results in a discrete probability distribution over the integers called the hypergeometric distribution.

Mathematica has a built–in function for the hypergeometric distribution,

`HypergeometricDistribution[` *arg1, arg2, arg3* **`]`**

demanding, as you might expect, three arguments.

The first argument is the number of draws n where here $n = 5$. The second argument is n_{succ}, the number of "successes," where a success is defined as the drawing of a red ball. Since there are $M = 7$ red balls in the urn, $n_{\text{succ}} = 7$. The third and final argument is n_{tot}, the total number of balls in the urn. Since there are N balls in the urn, $n_{\text{tot}} = 10$. Fill in these three arguments to arrive at,

`HypergeometricDistribution[` *n, nsucc, ntot* **`]`**

For our current example, the discrete distribution function is found through,

```
PDF[HypergeometricDistribution[5, 7, 10], k]
```

where k represents the $0 \leq k \leq n$, or, in other words, the above expression will return the probability for 0 through 5 red balls in five draws from the urn for any feasible value of k.

To find the probability of all six cases of the probability for zero through five red balls, evaluate,

```
Table[PDF[HypergeometricDistribution[5, 7, 10], k], {k, 0, 5}]
```

which returns the list of probabilities for $k = 0$ through $k = 5$ red balls,

$$\{0,\ 0,\ \tfrac{1}{12},\ \tfrac{5}{12},\ \tfrac{5}{12},\ \tfrac{1}{12}\}$$

From this result, we can read off from the fourth element in the returned list that the probability for three red balls and two white balls in any order in five draws from the urn is 5/12 as calculated in Supplemental Exercise 6.5.6.

It is impossible to have drawn no red balls or one red ball because that means either four or five white balls were drawn from the urn. But the problem specified that there were only a total of $W \equiv (N - M) = 3$ white balls in the urn to begin with. Notice as well the absolutely imperative condition that all six probabilities sum to 1.

Supplemental Exercise 6.5.8: Does Jaynes's combinatorial formula for the hypergeometric distribution correspond to *Mathematica*'s output?

Solution to Supplemental Exercise 6.5.8:

In Supplemental Exercise 6.5.6, we verified Jaynes's combinatorial formula for the hypergeometric distribution,

$$P(A \,|\, B) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}$$

where A is the statement "Exactly r red balls in n draws, in any order."

Thus,

$$
\begin{aligned}
P(A = 3 \text{ red balls and } 2 \text{ white balls} \,|\, \text{Bernoulli urn scenario}) &= \frac{\binom{7}{3}\binom{3}{2}}{\binom{10}{5}} \\
&= \frac{7 \times 6 \times 5 \times 3 \times 5 \times 4 \times 3 \times 2}{3 \times 2 \times 10 \times 9 \times 8 \times 7 \times 6} \\
&= \frac{5 \times 3 \times 2}{9 \times 8} \\
&= \frac{5}{12}
\end{aligned}
$$

in agreement with the *Mathematica* result for $k = 3$ in the previous exercise.

If we examine just what *Mathematica* outputs for,

```
PDF[HypergeometricDistribution[5, 7, 10], k]
```

we find $\frac{1}{252}$ **Binomial[3, 5 - k] Binomial[7, k]**. The $\frac{1}{252}$ term is coming from Jaynes's term in the denominator $\binom{N}{n} = \binom{10}{5} = 252$. The two combinatorial terms in Jaynes's numerator correspond to what we see in the *Mathematica* expression as,

$$\binom{3}{2} \times \binom{7}{3}$$

Supplemental Exercise 6.5.9: Code a perhaps more transparent form for the hypergeometric distribution.

Solution to Supplemental Exercise 6.5.9:

Based on the last exercise, it might be preferable to reorder the arguments to the hypergeometric distribution other than what *Mathematica* imposes on us in order to reflect Jaynes's combinatorial formula. To that end, therefore, create a new version of the hypergeometric distribution with,

```
BernoulliUrn[total_, totalRed_, draws_, red_] :=
 Binomial[totalRed, red] * Binomial[total - totalRed, draws - red]
                          / Binomial[total, draws]
```

with **Table[BernoulliUrn[10, 7, 5, r], {r, 0, 5}]** returning the correct probabilities for zero through five red balls in five draws from the urn.

Supplemental Exercise 6.5.10: Apply Bayes's Theorem for yet another numerical verification in our current example.

Solution to Supplemental Exercise 6.5.10:

Since Chapter Six of Volume I was an introduction to Bayes's Theorem, we might be wondering why it hasn't yet factored into Jaynes's discussion of sampling without replacement. It never really does attain a prominent place in the initial part of his explanation.

I will fill in the gap with this example. What is the probability of drawing a white ball on the fourth and fifth draws given that all red balls were drawn on the first three draws? In other words, calculate the conditional probability via Bayes's Theorem,

$$P(W_5\, W_4 \,|\, R_3\, R_2\, R_1\, B) = \frac{P(W_5\, W_4\, R_3\, R_2\, R_1 \,|\, B)}{P(R_1\, R_2\, R_3 \,|\, B)}$$

The general formula to calculate the numerator in Bayes's Theorem has already been worked out in Supplemental Exercise 6.5.5,

$$P(W_5\, W_4\, R_3\, R_2\, R_1 \,|\, B) = \frac{(N-M)!\,(N-n)!\,M!}{(N-M-w)!\,(M-r)!\,N!}$$

From the fundamental model B, the denominator must be,

$$P(R_1\, R_2\, R_3 \,|\, B) = \frac{M}{N} \times \frac{M-1}{N-1} \times \frac{M-2}{N-2}$$

Substitute the numerical values,

$$\begin{aligned} P(W_5\, W_4 \,|\, R_3\, R_2\, R_1\, B) &= \frac{\frac{(10-3)!\,(10-5)!\,7!}{(10-7-2)!\,(7-3)!\,10!}}{\frac{7}{10} \times \frac{6}{9} \times \frac{5}{8}} \\ &= \frac{3!\,5!\,7!}{1!\,4!\,10!} \times \frac{10 \times 9 \times 8}{7 \times 6 \times 5} \\ &= \frac{3! \times 5}{10 \times 9 \times 8} \times \frac{10 \times 9 \times 8}{7 \times 6 \times 5} \\ &= \frac{1}{7} \end{aligned}$$

which is the same as $\frac{3}{7} \times \frac{2}{6}$ which arises from Jaynes's statement B.

Supplemental Exercise 6.5.11: Is there anything of a conceptual nature that bothers you about the development so far?

Solution to Supplemental Exercise 6.5.11:

Here is a further conceptual sticking point that raises its ugly head. One gets the impression that this inferential problem will ultimately demand an application of

Bayes's Theorem. Uncritical thinking will tend towards framing the problem as one where *data* is involved. After all, balls are being plucked from the urn, and probabilities of later draws are being calculated conditioned on which balls were drawn earlier.

However, perhaps to your surprise, you come to the eventual realization that there are no data involved! It is, as Jaynes says, a "pre–data problem." What is the rationale for an assignment to the *joint probability distribution*?

For example, what does Jaynes say is the "correct" assignment to the joint probability $P(R_1\, W_2\, R_3\, W_4\, R_5 \,|\, B)$? As discovered in the previous exercises, Jaynes asserts his choice of a compelling rationale under a model B, "the Bernoulli urn scenario." Thus,

$$P(R_1\, W_2\, R_3\, W_4, R_5 \,|\, B) = \frac{(N-M)!\,(N-n)!\,M!}{(N-M-w)!\,(M-r)!\,N!}$$

The invocation of the **Product Rule** as the solution to the assignment of the joint probability can be deceiving. What, you ask yourself, is the new probability of drawing a red ball after a white ball and a red ball were drawn? You would naturally gravitate to thinking along the lines of: Given the data of red ball on the first draw and a white ball on the second draw, do I now use Bayes's Theorem to calculate the probability of a red on the third draw?

Here is the crucial conceptual anchor point. We have NOT moved beyond trying to find some legitimate assignment as a probability to a joint statement $R_3 \wedge W_2 \wedge R_1$ prior to any balls drawn whatsoever from the urn. When we wrote $P(A_3, \overline{A}_2, A_1 \,|\, \mathcal{M}_k)$ in a previous section, we did not condition on the existence of any data. No probability expression like,

$$P(A_{N+3}, \overline{A}_{N+2}, A_{N+1} \,|\, \mathcal{D}) \equiv P(A_{N+3}, \overline{A}_{N+2}, A_{N+1} \,|\, A_1, A_2, \cdots, A_N)$$

was written down. No data existed at that juncture.

The same applies here for the sampling without replacement scenario. The IP is trying to make a probability assignment to any *potential* succession of draws from the urn logically and probabilistically prior to any actual draws from the urn. Jaynes constantly emphasizes in this assignment procedure, and I think he is correct, that assigning probabilities as a state of knowledge should not be based solely on physical reasoning. As a consequence, when the **Product Rule** shows us a later draw conditioned on earlier draws, it becomes very confusing.

It is quite true that Jaynes said explicitly in his statement B that a ball was drawn, its color recorded, and the process repeated. And yet I get the impression from everything else he said afterwards that the probabilities to be calculated had nothing to do with this physical reality of how many balls were in the urn, how the balls were drawn, the physical construction of the balls and the urn, and so on, but rather everything to do with, as he said, any *logical connections* as might be imposed through some model.

Here is one way to make a vivid distinction. Borrow a technical term from English grammar, and think about this kind of inferential scenario of sampling without replacement as entering the *subjunctive mood*. Generically, the language used by the subjunctive mood looks something like: "*If* such and such *were* to be the case, then reality *would be* different" as in "If I were rich, then I wouldn't have to write books for a living."

So, it seems to me, upon entering the subjunctive mood, the IP really is asking itself the question: What might the probability of a red ball be on the third draw if it were true that a white ball and a red ball had happened to be drawn on the first two draws?

What are the implications for the solidus in our probability expressions? In the past, we have used phrases like, "given that," "assuming that," "conditioned on," "W_2 and R_1 actually happened," synonymously and interchangeably. Do we now have to add the qualifying phrase for statements appearing to the right of the solidus, "In a subjunctive frame of mind, hypothetically consider the fact that W_2 and R_1 might have happened."?

This whole drawing from an urn scenario is beginning to take on the appearance, contrary to what we were promised as an uncomplicated foundation on which to construct further complicated inferential scenarios, of bogging down into, what must be admitted, something not so straightforward after all.

I repeat my plaintive question: Why drove Jaynes to place this material so early on in his introduction to probability theory?

Supplemental Exercise 6.5.12: Prior to these exercises, where else has the hypergeometric distribution been discussed?

Solution to Supplemental Exercise 6.5.12:

Those who have read my Volume III may remember that I spent a considerable amount of time recapitulating Jeffreys's objections to the Bayes–Laplace uniform prior over model space. One of the precursor concepts leading eventually over many years to the development of his "Jeffreys's prior" was, in fact, sampling from a population without replacement.

During his investigation into sampling without replacement, Jeffreys derived the same formula as Jaynes for the hypergeometric distribution, albeit, not surprisingly, in his typically awkward notation. Of course, the attribution should be reversed; Jeffreys has priority. There is some merit in comparing Jeffreys's development of sampling without replacement, also leading to the hypergeometric distribution, with the just discussed version as presented by Jaynes.

Supplemental Exercise 6.5.13: Provide an alternative version of Jaynes's formula for the hypergeometric distribution in order to compare it with Jeffreys's version.

Solution to Supplemental Exercise 6.5.13:

On page 68 of his book, Jaynes provides us with a more intuitive formula for the hypergeometric distribution,

$$h(r) = \frac{\binom{R}{r}\binom{W}{w}}{\binom{R+W}{r+w}}$$

Calibrating with the notation in his other formula,

$$R = M$$

$$W = N - M$$

$$R + W = N$$

$$r + w = n$$

In this format, it can be more readily compared to Jeffreys's version [14, pg. 59],

$$P(l, m \,|\, H) = {}^{r}C_{l}\ {}^{s}C_{m} \,/\, {}^{r+s}C_{l+m}$$

To begin the translation, convert Jeffreys's idiosyncratic combinatorial notation of ${}^{x}C_{y}$ into,

$$P(l, m \,|\, H) = \frac{\binom{r}{l}\binom{s}{m}}{\binom{r+s}{l+m}}$$

With this much accomplished, it is much easier to see that in Jeffreys's notation, $r \equiv R$, the total number of red balls in the urn, $l \equiv r$, the number of red balls drawn, $s \equiv W$, the total number of white balls in the urn, $m \equiv w$, the number of white balls drawn from the urn, and $r + s \equiv R + W$, the total number of red and white balls in the urn, $l + m \equiv r + w = n$, the sample size.

Jeffreys's version comes up with the same probability for our running numerical example of drawing three red balls and two white balls as a sample size of five,

$$P(l = 3, m = 2 \,|\, H) = \frac{\binom{7}{3}\binom{3}{2}}{\binom{7+3}{3+2}} = \frac{5}{12}$$

Supplemental Exercise 6.5.14: Compare Jeffreys's rationale for the form of the hypergeometric distribution to Jaynes's recently examined version.

Solution to Supplemental Exercise 6.5.14:

Jeffreys chooses to characterize our current inferential scenario of drawing balls from an urn as sampling from a population of two types. He poses the question: What

is the probability of sampling l individuals of Type I and m individuals of Type II from a known population consisting of r Type Is and s Type IIs?

He also calls this a "direct probability" and "sampling without replacement" but he relies exclusively on combinatorial arguments. He does not use Jaynes's rationale of probabilities changing with each draw from the urn. Jeffreys says that considering the sample size taken from the population, each set is "equally likely" to be taken. Thus, returning to the Bernoulli urn scenario, any sample of size $n = 5$ taken from the $N = 10$ balls in the urn has an equal probability of,

$$\frac{1}{\binom{N}{n}} \equiv \frac{1}{\binom{R+W}{r+w}} \equiv \frac{1}{\binom{r+s}{l+m}} = \frac{1}{252}$$

So whether we are considering the probability of, say,

$$P(R_1\, R_2\, R_3\, R_4\, R_5 \,|\, B) \text{ or } P(R_1\, W_2\, W_3\, R_4\, W_5 \,|\, B)$$

they both have the same probability of 1/252.

Again, relying strictly on a combinatorial argument, Jeffreys explains that there are $\binom{r}{l}$ different ways to obtain the l red balls multiplied by $\binom{s}{m}$ different ways to obtain the m white balls. This number counts up all of the ways for the specified l and m, and thus this value is the numerator that goes over the denominator of 252.

The following numerical example where the numbers are adjusted to make it easy to spell everything out in exquisite detail should make this clear. Suppose that the entire population consists of a total of $N = 6$ individuals and each person has a name. The population is divided into two types, $r = 4$ of Type I, "The Ready" (R), and $s = 2$ of Type II, "The Willing" (W). What is the probability of sampling $l = 2$ "Ready" individuals and $m = 1$ "Willing" individual from a sample of size $n = 3$ drawn from the population?

First, there must be $\binom{6}{3} = 20$ sets of samples of size three. As part of that exquisite detail, here are the names of all six individuals: the four Ready individuals are 1) Rick, 2) Rachel, 3) Robert, and 4) Rita, and the two Willing individuals are 1) Wilma, and 2) Wycliffe.

For example, one of these 20 samples might consist of Robert, Rita, and Wycliffe. Notationally write this sample as $R^{(3)}\, R^{(4)}\, W^{(2)}$. This particular sample and all of the other samples have a probability of 1/20. We want to maintain a distinction in the notation between Jaynes's $R_1\, W_2\, R_3$ where the *subscript* indicates the trial on which a ball was drawn, and Jeffreys's sample drawn "all at once" where the *superscript* indicates an individual's name.

See Figure 6.3 at the top of the next page for the complete listing of all 20 samples of size 3. The first block starting on the left lists the,

$$\binom{R}{r}\binom{W}{w} \equiv \binom{r}{l}\binom{s}{m} \equiv \binom{4}{3}\binom{2}{0} = 4$$

3 Ready	2 Ready 1 Willing	2 Ready 1 Willing	1 Ready 2 Willing
$R^{(1)}\ R^{(2)}\ R^{(3)}$	$R^{(1)}\ R^{(2)}\ W^{(1)}$	$R^{(1)}\ R^{(2)}\ W^{(2)}$	$R^{(1)}\ W^{(1)}\ W^{(2)}$
$R^{(1)}\ R^{(2)}\ R^{(4)}$	$R^{(1)}\ R^{(3)}\ W^{(1)}$	$R^{(1)}\ R^{(3)}\ W^{(2)}$	$R^{(2)}\ W^{(1)}\ W^{(2)}$
$R^{(1)}\ R^{(3)}\ R^{(4)}$	$R^{(1)}\ R^{(4)}\ W^{(1)}$	$R^{(1)}\ R^{(4)}\ W^{(2)}$ *	$R^{(3)}\ W^{(1)}\ W^{(2)}$
$R^{(2)}\ R^{(3)}\ R^{(4)}$	$R^{(2)}\ R^{(3)}\ W^{(1)}$	$R^{(2)}\ R^{(3)}\ W^{(2)}$	$R^{(4)}\ W^{(1)}\ W^{(2)}$
	$R^{(2)}\ R^{(4)}\ W^{(1)}$	$R^{(2)}\ R^{(4)}\ W^{(2)}$	
	$R^{(3)}\ R^{(4)}\ W^{(1)}$	$R^{(3)}\ R^{(4)}\ W^{(2)}$	
4/20	6/20	6/20	4/20
1/5	3/5		1/5

Figure 6.3: *All twenty possible samples of size 3 from a population of six individuals. The superscript number indicates a named individual.*

four sets of all three Ready individuals. The next two blocks over list the first six sets and second six sets from the total of,

$$\binom{R}{r}\binom{W}{w} \equiv \binom{r}{l}\binom{s}{m} \equiv \binom{4}{2}\binom{2}{1} = 12$$

twelve sets of two Ready individuals and one Willing individual. The last block on the right lists the,

$$\binom{R}{r}\binom{W}{w} \equiv \binom{r}{l}\binom{s}{m} \equiv \binom{4}{1}\binom{2}{2} = 4$$

four sets of one Ready individual and two Willing individuals. The number in each block divided by the total number is shown below each of the four blocks.

For example, in the third block over, the third row marked by an asterisk, shows a possible set of size three consisting of Rick, Rita, and Wycliffe sampled from the entire population. This is just one of the 12 possibilities of two Ready individuals and one Willing individual.

The hypergeometric distribution for $k = 0$ through $k = 3$ Ready individuals would be found as listed in Table 6.2.

Table 6.2: *The probabilities for the number of Ready individuals in a sample of size $n = 3$ taken from a population of $N = 6$.*

Ready individuals			
0	1	2	3
Probability			
0	1/5	3/5	1/5

According to Jeffreys, each one of these 20 possibilities has the same probability of 1/20. The probability of $l = 2$ Ready individuals and $m = 1$ Willing individual, without regard to their names, is found by summing over all 12 possibilities. Thus,

$$P(l = 2, m = 1 \mid H) = 12/20 = 3/5$$

Have *Mathematica* verify the probabilities of the k Ready individuals shown in the table through an evaluation of the hypergeometric distribution,

```
Table[PDF[HypergeometricDistribution[3, 4, 6], k], {k, 0, 3}]
```

The results from *Mathematica* confirm the probabilities shown in the table.

6.6 Non–binary Variables

In section 6.4 of Volume I, we broached the issue of generalizing a statement away from the two categories of TRUE or FALSE with the introduction of the Shakespeare example. Recall that statement A was "Shakespeare wrote the plays attributed to him." Therefore, the notation $\overline{A}$ was interpreted as, "It is FALSE that Shakespeare wrote the plays attributed to him." But statement $\overline{A}$ was then broken down further into another statement $A_\star$, "Marlowe wrote the plays attributed to Shakespeare." with then $\overline{A}_\star$, "Marlowe did not write the plays attributed to Shakespeare."

To finish up, we attached the meaning of "de Vere wrote the plays attributed to Shakespeare." to $\overline{A}_\star$. But an argument might be made that $\overline{A}_\star$ really implies "de Vere or *anyone else* other than Shakespeare or Marlowe wrote the plays."

One can continue to rely on *Mathematica*'s Boolean related built–in functions when dealing with variables that are no longer binary. In the Supplemental Exercise below, we consider an inference involving two statements A and C where statement A can be measured at four different values. Statement C remains a binary variable. This scenario would lead to a state space of dimension $n = 8$, with eight cells in the joint probability table.

A Boolean function consisting of three binary variables A, B, and C would also lead to a state space of dimension $n = 8$. We can match up the four combinations of A and B with the four possible measurements of A. All of *Mathematica*'s built–in Boolean functions could then be applied to this three variable Boolean function. Suitable care would have to be taken in matching the eight cells of the original non–binary A and C joint probability table with how *Mathematica* orders things.

Supplemental Exercise 6.6.1: Compute the conditional probability of obtaining the third measurement value for statement A given knowledge that C was measured at its second value.

Solution to Supplemental Exercise 6.6.1:

Using Bayes's Theorem for this inference, we have,

$$P(A = a_3 \,|\, C = c_2, \mathcal{M}_k) = \frac{P(A = a_3, C = c_2 \,|\, \mathcal{M}_k)}{P(C = c_2 \,|\, \mathcal{M}_k)}$$

$$= \frac{P(A = a_3, C = c_2 \,|\, \mathcal{M}_k)}{\sum_{j=1}^{4} P(A = a_j, C = c_2 \,|\, \mathcal{M}_k)}$$

The numerical assignment of probabilities to the joint probability table must be conditioned on the information provided through model $\mathcal{M}_k$. Let's take Rule 110 as something already familiar to us as a three variable logic function that can inspire a large class of models.

So, as before, we can construct the Boolean function,

```
g = BooleanFunction[110, 3]
```

followed by,

```
BooleanTable[g[a, b, c]]
```

to find the locations in the joint probability table where 0s must be placed.

See Figure 6.4 below for a summary of this output fashioned to get ready for the assignments in an eight cell joint probability table for non–binary variable A and C.

T, T, T	T, T, F	T, F, T	T, F, F	F, T, T	F, T, F	F, F, T	F, F, F
A, B, C	A, B, $\overline{C}$	A, $\overline{B}$, C	A, $\overline{B}$, $\overline{C}$	$\overline{A}$, B, C	$\overline{A}$, B, $\overline{C}$	$\overline{A}$, $\overline{B}$, C	$\overline{A}$, $\overline{B}$, $\overline{C}$
Cell 1	Cell 5	Cell 2	Cell 6	Cell 3	Cell 7	Cell 4	Cell 8
F	T	T	F	T	T	T	F
0	q	q	0	q	q	q	0

Figure 6.4: *Using the* Mathematica *Boolean operators on three binary statements to help construct a joint probability table for a non–binary variable.*

See Figure 6.5 below for the actual eight cell joint probability table.

	A		$\overline{A}$	
	B	$\overline{B}$	B	$\overline{B}$
C	1: ABC, 0, $a_1 c_1$	2: $A\overline{B}C$, q, $a_2 c_1$	3: $\overline{A}BC$, q, $a_3 c_1$	4: $\overline{A}\overline{B}C$, q, $a_4 c_1$
$\overline{C}$	5: $AB\overline{C}$, q, $a_1 c_2$	6: $A\overline{B}\overline{C}$, 0, $a_2 c_2$	7: $\overline{A}B\overline{C}$, q, $a_3 c_2$	8: $\overline{A}\overline{B}\overline{C}$, 0, $a_4 c_2$

Figure 6.5: *The final* 2×4 *joint probability table ensuing from Figure 6.4.*

We will match the two binary variables A and B with the four measurements of A as,

$$AB \to a_1$$

$$A\overline{B} \to a_2$$

$$\overline{A}B \to a_3$$

$$\overline{A}\,\overline{B} \to a_4$$

There must be probability assignments of 0 in cells 1, 6, and 8 from this model. Suppose we choose the common assignment of q for the remaining five cells. After substituting these numerical values into Bayes's Theorem, we can compute the probability of seeing the third possible measurement of A as 1/2,

$$\begin{aligned} P(A = a_3 \,|\, C = c_2, \mathcal{M}_k) &= \frac{P(A = a_3, C = c_2 \,|\, \mathcal{M}_k)}{\sum_{i=1}^{4} P(A = a_i, C = c_2 \,|\, \mathcal{M}_k)} \\ &= \frac{\text{Cell 7}}{\text{Cell 5} + \text{Cell 6} + \text{Cell 7} + \text{Cell 8}} \\ &= \frac{q}{q + 0 + q + 0} \\ &= 1/2 \end{aligned}$$

There is some subtlety here as one must match up the *Mathematica* outputs for **`Tuples[{T, F}, 3]`** and the layout of the joint probability table, which we were careful to do in Figure 6.4.

See Figure 6.6 below to make sure that the assignments from Figure 6.4 get translated properly over into the joint probability table of Figure 6.5. Notice in particular that the joint probability table is constructed differently than the way I usually present it.

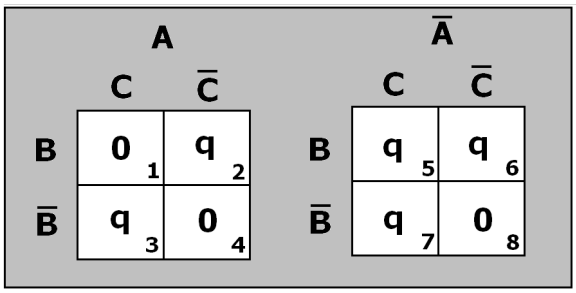

Figure 6.6: *A joint probability table constructed according to the ordering dictated by* **`Tuples[{T, F}, 3]`**.

Chapter 7

Generalizing Logic with Probability

7.1 Revisiting the Logic Puzzle from Chapter Two

One of the major goals of our effort is to substantiate the claim that inference using probability theory generalizes deduction using Classical Logic. Therefore, probability theory must be able to reproduce the same results as provided through the formal manipulations of Boolean Algebra when applied to logic. In this case, an inference must return probabilities of either 1 or 0 corresponding to a T or F.

In Chapter Seven of Volume I, we looked at the classical syllogisms of *modus ponens*, *modus tollens*, the *Process of Elimination*, and *Proof by Cases*. We were able to verify that inferencing using probability theory, by returning 1 or 0 as probabilities, reproduced the correct answers obtained by deduction using logic.

Let's revisit the second logic puzzle as it was introduced in Chapter Two of these Supplemental Exercises. There were indeed some puzzling aspects encountered during the course of attempting to confirm Mendelson's results. The most troubling outcome was that a tautology did not exist, and yet it seemed that the conclusion was always true!

The formal manipulation rules of Boolean Algebra carried out in the abstract will always leave one wondering: what just happened? I suggest that examining these logic puzzles from the probabilistic viewpoint is much less confusing. To advance that objective, we will see what happens when we attempt to leverage Bayes's Theorem on this puzzle. A very interesting moral is revealed that impacts our decision as to when Bayes's Theorem can be legitimately employed.

Supplemental Exercise 7.1.1: Present the solution to the second logic puzzle of Chapter Two as an inferential problem using probability.

Solution to Supplemental Exercise 7.1.1:

Use Bayes's Theorem to find the probability for R, the statement, "Taxes will be raised." conditioned on the truth of statements B and P. Statement B is, "The budget will be cut." and statement P is, "Prices remain stable." Mendelson said that the premises logically implied the conclusion that $\overline{R}$, "Taxes will not be raised." He claimed that the argument was a correct one.

Our analysis in Chapter Two, relying as it did solely on Boolean Algebra, and ignoring probability theory, was, however, not able to confirm that a tautology existed. There was one setting of the variables that evaluated to F for the premises, but evaluated to T for the conclusion. Thus, there was no logical equivalency across all eight cases. There was no tautology. Can probability and inferencing shed light on this confusion?

If probability theory as a generalization of Classical Logic were to capture Mendelson's argument, then the probability of R should equal 0 in all cases for B and P. Then, as Mendelson would like to see, the probability that taxes will NOT be raised is a certainty, $[\,P(R = F) = 1\,] \equiv [\,P(R = T) = 0\,]$. There are four possible cases, and the probability for $R = T$ is indeed 0 for three of these cases as we now show.

Write the conditional probability expression on the left hand side of Bayes's Theorem for the first case as,

$$P(R = T \mid B = T, P = T, \mathcal{M}_k)$$

The *information* in model $\mathcal{M}_k$ is the information provided by writing out the symbolic expression in the premises,

$$\mathcal{M}_k \equiv (\overline{B} \rightarrow (P \leftrightarrow R)) \wedge (R \rightarrow \overline{B}) \wedge (P \rightarrow \overline{R})$$

The DNF for this model will tell us where any 0s are located in the joint probability table. Refer back to Supplemental Exercise 2.2.8. Since there were five Fs in the DNF for the premises, there will be five 0s located at the appropriate cell of the joint probability table. See Figure 7.1 for the location of the five 0s in the eight cell joint probability table. Because it makes no difference to our goal here, the three positive probabilities are set to 1/3.

With the joint probability table now determined by conditioning on the truth of the premises, it is easy to calculate the probability for R for all four cases. Insert 0s and 1/3 for the joint probabilities in Bayes's Theorem by consulting the appropriate cell of the joint probability table.

	B				$\overline{B}$			
	P	$\overline{P}$			P	$\overline{P}$		
R	0 (1)	0 (2)	0	R	0 (5)	0 (6)	0	0
$\overline{R}$	1/3 (3)	1/3 (4)	2/3	$\overline{R}$	0 (7)	1/3 (8)	1/3	1
	1/3	1/3	2/3		0	1/3	1/3	
					1/3	2/3		1

Figure 7.1: *The joint probability table for the second logic puzzle.*

$$P(R\,|\,B,P,\mathcal{M}_k) = \frac{P(B,P,R\,|\,\mathcal{M}_k)}{P(B,P,R\,|\,\mathcal{M}_k) + P(B,P,\overline{R}\,|\,\mathcal{M}_k)}$$

$$= \frac{0}{0+1/3} = 0$$

$$P(R\,|\,\overline{B},P,\mathcal{M}_k) = \frac{P(\overline{B},P,R\,|\,\mathcal{M}_k)}{P(\overline{B},P,R\,|\,\mathcal{M}_k) + P(\overline{B},P,\overline{R}\,|\,\mathcal{M}_k)}$$

$$= \frac{0}{0+0} = \textbf{Undefined !}$$

$$P(R\,|\,B,\overline{P},\mathcal{M}_k) = \frac{P(B,\overline{P},R\,|\,\mathcal{M}_k)}{P(B,\overline{P},R\,|\,\mathcal{M}_k) + P(B,\overline{P},\overline{R}\,|\,\mathcal{M}_k)}$$

$$= \frac{0}{0+1/3} = 0$$

$$P(R\,|\,\overline{B},\overline{P},\mathcal{M}_k) = \frac{P(\overline{B},\overline{P},R\,|\,\mathcal{M}_k)}{P(\overline{B},\overline{P},R\,|\,\mathcal{M}_k) + P(\overline{B},\overline{P},\overline{R}\,|\,\mathcal{M}_k)}$$

$$= \frac{0}{0+1/3} = 0$$

The problematical conditions for B, P, and R involve the second application of Bayes's Theorem as they were presented above. This problematical case is the exactly the one that our previous analysis in Chapter Two identified as violating the tautology, that is, $B = F$, $P = T$, and $R = F$. Clearly, strictly from the arithmetic alone in Bayes's Theorem, we have a problem in the division by 0.

It will never happen that the assignments $B = F$ and $P = T$ might be TRUE as asserted to the right of the conditioned upon symbol under the information from the model $\mathcal{M}_k$ of the premisses.

Moreover, from the standpoint of the formal rules, this second condition using Bayes's Theorem is conditioning on the truth of $\overline{B}P$. But, an application of the **Sum Rule** shows that,

$$P(\overline{B}P \,|\, \mathcal{M}_k) = P(\overline{B}PR \,|\, \mathcal{M}_k) + P(\overline{B}P\overline{R} \,|\, \mathcal{M}_k) = 0$$

Probability theory has indeed helped us to understand what went wrong here. The answer resides in this infrequently mentioned requirement for Bayes's Theorem. Whatever appears as the statement to the right of the conditioned upon symbol is assumed to be true. Bayes's Theorem is not allowed to condition on a statement that is assumed true, but under the information in the model **can never be true!**

The marginal probability of $B = F$ and $P = T$ over both conditions for R can never happen; it has a probability of 0 under the information provided by the premises! Bayes's Theorem is therefore undefined.

Supplemental Exercise 7.1.2: What immediate solution suggests itself?

Solution to Supplemental Exercise 7.1.2:

Assign a ridiculously low numerical assignment to the problematic cell like,

$$P(\overline{B}P\overline{R} \,|\, \mathcal{M}_k) = 10^{-100}$$

Of course, this assignment must still be a positive number whose value can be subtracted from any one of the other assignments. See section 7.4 of these Supplemental Exercises for further commentary on this resolution of problematic 0s.

7.2 *Mathematica* and Some Classical Syllogisms

The reference manual for the Wolfram Language provides examples of how to prove a few classical syllogisms. *Mathematica* employs the tactic of proving them through application of the familiar built–in function **`TautologyQ[ ]`** for Boolean functions. These same syllogisms were introduced in Chapter Seven of Volume I.

The first two examples given are *modus ponens* and *modus tollens*. Then, two more syllogisms are presented called *modus tollendo ponens* and *modus ponendo tollens* (time to brush up on that rusty Latin!). Finally, *reductio ad absurdum* is demonstrated as correct through **`TautologyQ[ ]`**.

Of course, my intent in Chapter Seven was to introduce the generalization of logic through probability arguments. So, I proved these syllogisms by demonstrating that probabilities as calculated by Bayes's Theorem resulted in 0 or 1.

The symbolic expressions for the syllogisms are shown in the same notation as presented in the reference manual. The expression on the right hand side of the final `IMPLIES` ($\Rightarrow$) was the statement calculated as having a probability of 0 or 1.

Cases 3 and 4 below, labeled as *modus tollendo ponens* and *modus ponendo tollens* were, in my terminology, examples of the *process of elimination*. Exercise 7.9.4 treated the *reductio ad absurdum* syllogism from the viewpoint of a probability calculation which must return a value of 0 or 1.

1. ***modus ponens*** $((a \Rightarrow b) \wedge a) \Rightarrow b$

```
TautologyQ[Implies[And[Implies[a, b], a], b]]
```

$\boldsymbol{P(B \mid A, A \rightarrow B) = 1}$

2. ***modus tollens*** $((a \Rightarrow b) \wedge \neg b) \Rightarrow \neg a$

```
TautologyQ[Implies[And[Implies[a, b], Not[b]], Not[a]]]
```

$\boldsymbol{P(A \mid \overline{B}, A \rightarrow B) = 0}$

3. ***modus tollendo ponens*** $((a \vee b) \wedge \neg a) \Rightarrow b$

```
TautologyQ[Implies[And[Or[a, b], Not[a]], b]]
```

$\boldsymbol{P(B \mid \overline{A}, A \vee B) = 1}$

4. ***modus ponendo tollens*** $(\neg(a \wedge b) \wedge a) \Rightarrow \neg b$

```
TautologyQ[Implies[And[Not[And[a, b]], a], Not[b]]]
```

$\boldsymbol{P(B \mid A, [\,\overline{A \wedge B}\,]) = 0}$

5. ***reductio ad absurdum*** $((a \Rightarrow b) \wedge (a \Rightarrow \neg b) \Rightarrow \neg a$

```
TautologyQ[Implies[And[Implies[a, b], Implies[a, Not[b]]],
            Not[a]]]
```

$\boldsymbol{P(A \mid [\,(A \rightarrow B) \wedge (A \rightarrow \overline{B})\,]) = 0}$

I guess it goes without saying that *Mathematica*'s tactic to prove these classical syllogisms correct by evaluating **`TautologyQ[ ]`** returned **`True`** in all cases.

Supplemental Exercise 7.2.1: Exercise 7.9.4 of Volume I actually worked on something different than the above Case 5 *reductio ad absurdum*.

Solution to Supplemental Exercise 7.2.1:

In Exercise 7.9.4, I showed that the probability of statement A was 1 when two models implemented the information in the logic expression $(A \rightarrow B) \wedge (\overline{A} \rightarrow \overline{B})$ and its negation. This was my probabilistic proof of *reductio ad absurdum*. The Wolfram Language reference manual used the mirror image version for the two models, namely, $(A \rightarrow B) \wedge (A \rightarrow \overline{B})$ and its negation to prove that A must be FALSE. Here is the equivalent probabilistic treatment.

We now let the statement Z equal the model for *Mathematica* version of the *reductio ad absurdum* syllogism,

$$Z \equiv (A \rightarrow B) \wedge (A \rightarrow \overline{B})$$

All we have to do is show that the numerator in Bayes's Theorem,

$$P(A\,|\,Z) = \frac{P(AZ)}{P(Z)}$$

is equal to 0 to prove that $P(A\,|\,Z) = 0$, or, in other words, that A must be FALSE. The numerator is the joint probability that statement A and the *reductio ad absurdum* logic expression are both TRUE.

Carrying out the required Boolean operations within the probability operator, we have,

$$\begin{aligned}
P(A\,|\,Z) &= \frac{P(AZ)}{P(Z)} \\
P(AZ) &= P(A \wedge [\,(A \to B) \wedge (A \to \overline{B})\,]\,) \\
&= P(A \wedge [\,(\overline{A} \vee B) \wedge (\overline{A} \vee \overline{B})\,]\,) \\
&= P(A \wedge [\,\overline{A}\,\overline{A} \vee \overline{A}\,\overline{B} \vee B\overline{A} \vee B\overline{B}\,]\,) \\
&= P(\,[\,A\overline{A}\,\overline{A} \vee A\overline{A}\,\overline{B} \vee AB\overline{A} \vee AB\overline{B}\,]\,) \\
&= P(F) \\
P(A\,|\,Z) &= 0
\end{aligned}$$

Supplemental Exercise 7.2.2: Have *Mathematica* verify the correctness of the Boolean operations carried out in the previous exercise.

Solution to Supplemental Exercise 7.2.2:

Use the built–function **`Distribute[`***expr, g***`]`** to confirm that,

$$P(A \wedge [\,(\overline{A} \vee B) \wedge (\overline{A} \vee \overline{B})\,]\,) = P(\,[\,A\overline{A}\,\overline{A} \vee A\overline{A}\,\overline{B} \vee AB\overline{A} \vee AB\overline{B}\,]\,)$$

```
Distribute[And[Or[Not[a], b], Or[Not[a], Not[b]], a],
                                                Or] // FullForm
```

returns the following expression confirming our derivation in the last exercise,

```
Or[And[Not[a], Not[a], a], And[Not[a], Not[b], a],
   And[b, Not[a], a], And[b, Not[b], a]]
```

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

The **`Distribute[ ]`** function was initially discussed together with an example in **Appendix B** of Volume I.

This Boolean expression returned by **Distribute[ ]** does indeed evaluate to FALSE through both,

```
TautologyQ[Equivalent[Distribute[And[Or[Not[a],b],Or[Not[a],
             Not[b]],a],Or],False]]
```

and $A\overline{A}\,\overline{A} \vee A\overline{A}B \vee AB\overline{A} \vee AB\overline{B} \leftrightarrow F$,

```
TautologyQ[Equivalent[Or[And[a,Not[a],Not[a]],
   And[a,Not[a],Not[b]],And[a,b,Not[a]],And[a,b,Not[b]]],
                                                  False]]
```

returning **True**.

7.3 Revisiting the Life on Mars Scenario

I illustrated how to avoid an invalid application of *modus ponens* by resorting to probability theory in section 7.6.3 of Volume I. This scenario involved an inference about the presence of life and water on Mars. Through Classical Logic and *modus ponens* we are allowed, if we accept the premise that the presence of life implies the presence of water, and the further premise that life is actually present, to reason to the conclusion that water is also present.

Expressed symbolically as a probability expression, $P(W \,|\, L, L \to W) = 1$. In other words, the IP can use probability theory to make an *inference* about which there is no doubt, mimicking exactly the *deductive* certainty of Classical Logic. But if an IP tries to use Classical Logic to reason about the presence of life given the presence of water on Mars, the deductive machinery grinds to a halt.

If the IP, in desperation, then turns to the appropriate inference for an answer, it finds that the expression $P(L \,|\, W, L \to W)$ will yield up a number somewhere between 0 and 1. In other words, an inference is something that *can* be computed. Furthermore, the output from an inference represents a quantitative measure for the degree of belief in the truth of the statement that life is present when conditioned on the truth of the statement that water is present. The IP can no longer count on certainty in the truth of the statement, as it could before, when deduction and logic were also valid. But at least it is not stymied and can provide a qualified answer.

The price to be paid for this fortuitous turn of events is that the IP must now entertain the notion that the original state space must now be enlarged to include statements about an entire space of models. The information in these models is what is going to allow numerical assignments of probability to be made to the joint statements in the newly enlarged state space.

Although we are restricting the class of models to the logic function `IMPLIES`, this still permits a potentially huge number of models. In the upcoming supplemental exercise, we will examine the computations involved in an inference when the IP

chooses $\mathcal{M} = 3$ "plausible" models. All three models will adhere nevertheless to the fundamental *sine qua non* of a logical implication like $L \rightarrow W$. Implication means that L and $\overline{W}$ must be FALSE. Translating this requirement into assigned numerical values in a joint probability table means that the cell in the table that indexes the joint statement $L \wedge \overline{W}$ must have a probability assignment of 0.

Probability theory allows us the freedom either to set up three separate four cell joint probability tables, three tables for each separate model, or to construct one large twelve cell table with the models already folded in. For this exercise, we shall examine the latter option.

Supplemental Exercise 7.3.1: Show a twelve cell joint probability table for the Life on Mars scenario.

Solution to Supplemental Exercise 7.3.1:

We mentioned in the introductory remarks that, solely for expository purposes, we shall restrict ourselves to just three models, each an "advocate" for a particular attitude about the scientific credibility of life and water on Mars. It is a given that the cell for $L \wedge \overline{W}$ is always going to be assigned a 0 for all three models. For that matter, that cell would always be assigned a 0 for any model from the class of models implementing $L \rightarrow W$.

Each cell in the joint probability table shown below in Figure 7.2 assigns a numerical probability for a joint statement, $P(L, W, \mathcal{M}_1)$ through $P(\overline{L}, \overline{W}, \mathcal{M}_3)$.

$\mathcal{M}_1$	L	$\overline{L}$		$\mathcal{M}_2$	L	$\overline{L}$		$\mathcal{M}_3$	L	$\overline{L}$		
W	1/9 (1)	1/9 (2)	2/9	W	1/27 (5)	1/27 (6)	2/27	W	1/6 (9)	1/18 (10)	2/9	14/27
$\overline{W}$	0 (3)	1/9 (4)	1/9	$\overline{W}$	0 (7)	7/27 (8)	7/27	$\overline{W}$	0 (11)	1/9 (12)	1/9	13/27
	1/9	2/9	1/3		1/27	8/27	1/3		1/6	1/6	1/3	
									17/54	37/54		1

Figure 7.2: *A joint probability table involving three models for the Life on Mars scenario following from $L \rightarrow W$.*

The first model $\mathcal{M}_1$ takes the nonchalant view that the three remaining joint statements might all be equally true. So it assigns 1/9 to cells 1, 2, and 4. The second model $\mathcal{M}_2$ is an advocate for sterility on Mars. The highest probability by far is assigned to $\overline{L} \wedge \overline{W}$, although the other two statements are not completely ruled out. The third and final model $\mathcal{M}_3$ is rather more disposed to finding both life and water, although not overwhelmingly so. Thus, the highest probability is assigned to $L \wedge W$, with slightly lesser probabilities assigned to $\overline{L} \wedge W$ and $\overline{L} \wedge \overline{W}$.

Notice particularly that the marginal sum for each model results in,

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = P(\mathcal{M}_3) = 1/3$$

This choice for the prior probability means that the IP is "totally uninformed" about what actually is the ground truth about water and life on Mars. Obviously, the universal constraint that all assigned probabilities sum to 1 must be satisfied. The marginal sums for the probabilities of life $P(L) = 17/54$ and probability of water $P(W) = 14/27$ can be checked as well.

Supplemental Exercise 7.3.2: Show the derivation that arrives at the probability of life on Mars given the acknowledged presence of water.

Solution to Supplemental Exercise 7.3.2:

If we went to all the trouble of constructing a twelve cell joint probability table, then a generic joint probability $P(L, W, \mathcal{M}_k)$ must mean something important. By the **Product Rule**, the joint probability can be decomposed as,

$$P(L, W, \mathcal{M}_k) = P(L \,|\, W, \mathcal{M}_k) \times P(W \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

By the **Sum Rule**, we marginalize over the models,

$$P(L, W) = \sum_{k=1}^{3} P(L \,|\, W, \mathcal{M}_k)\, P(W \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

The inference that we want to make is by Bayes's Theorem,

$$P(L \,|\, W) = \frac{P(L, W)}{P(W)}$$

Thus, the right hand side for $P(L, W)$ must be divided by $P(W)$,

$$P(L \,|\, W) = \frac{\sum_{k=1}^{3} P(L \,|\, W, \mathcal{M}_k)\, P(W \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(W)}$$

But,

$$P(\mathcal{M}_k \,|\, W) = \frac{P(W \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(W)}$$

so we have,

$$P(L \,|\, W) = \sum_{k=1}^{3} P(L \,|\, W, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, W)$$

Supplemental Exercise 7.3.3: Carry out the computational details for finding the probability of life on Mars given the presence of water.

Solution to Supplemental Exercise 7.3.3:

Recognizing full well that we are making an inference as opposed to a deduction, we are now ready to compute a degree of belief in the truth of the statement that

life exists on Mars given that it is known that water is present. We were forbidden by Classical Logic from reaching any conclusion from the implication $L \to W$. Probability demurs on any such objection, and invites us to compute away.

We have already established that the IP decided that it was indifferent regarding the truth of the probability assignments made by each model. Therefore, a *prior* probability of $P(\mathcal{M}_k) = 1/3$ was assigned. However, *after* it becomes a fact that water is present, the IP must update its state of knowledge about all three models. This is the calculation of the *posterior* probability $P(\mathcal{M}_k \,|\, W)$ over model space, the second term on the right hand side of $P(L \,|\, W)$,

$$P(\mathcal{M}_k \,|\, W) = \frac{P(W \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)}{\sum_{k=1}^{3} P(W \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)}$$

Now, for each specific model,

$$\begin{aligned}
P(\mathcal{M}_1 \,|\, W) &= \frac{P(W \,|\, \mathcal{M}_1) \times P(\mathcal{M}_1)}{\sum_{k=1}^{3} P(W \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)} \\
P(W \,|\, \mathcal{M}_1) \times P(\mathcal{M}_1) &= P(W, \mathcal{M}_1) \\
&= 1/9 + 1/9 \\
P(W \,|\, \mathcal{M}_2) \times P(\mathcal{M}_2) &= P(W, \mathcal{M}_2) \\
&= 1/27 + 1/27 \\
P(W \,|\, \mathcal{M}_3) \times P(\mathcal{M}_3) &= P(W, \mathcal{M}_3) \\
&= 1/6 + 1/18 \\
\sum_{k=1}^{3} P(W \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) &= 2/9 + 2/27 + 12/54 \\
&= 6/27 + 2/27 + 6/27 \\
&= 14/27 \\
P(\mathcal{M}_1 \,|\, W) &= \frac{6/27}{14/27} = 3/7 \\
P(\mathcal{M}_2 \,|\, W) &= \frac{2/27}{14/27} = 1/7 \\
P(\mathcal{M}_3 \,|\, W) &= \frac{6/27}{14/27} = 3/7
\end{aligned}$$

The rules of probability theory have readjusted the IP's state of knowledge from a prior probability of 1/3 for each model to 3/7 for models $\mathcal{M}_1$ and $\mathcal{M}_3$ and 1/7 for model $\mathcal{M}_2$. All of this makes intuitive sense since these two models were more favorably disposed to the presence of water from the outset.

Whatever posterior probabilities for the models fall out from the computation must absolutely obey the universal constraint function of adding up to 1. It doesn't make any difference to the universal constraint function what the probability for the models, or for that matter, what the probabilities for any statement, might be conditioned on. They must still sum to 1 as our newly readjusted posterior probabilities certainly do satisfy.

With this much accomplished, the IP can now proceed to the first term in the probability for Life given the presence of Water,

$$P(L \mid W) = \sum_{k=1}^{3} P(L \mid W, \mathcal{M}_k)\, P(\mathcal{M}_k \mid W)$$

To fill in the first term for each of the three models refer back to Figure 7.2,

$$\begin{aligned}
P(L \mid W, \mathcal{M}_1) &= \frac{P(L, W \mid \mathcal{M}_1)}{P(W \mid \mathcal{M}_1)} \\
&= \frac{1/9}{1/9 + 1/9} \\
&= 1/2 \\
P(L \mid W, \mathcal{M}_2) &= \frac{P(L, W \mid \mathcal{M}_2)}{P(W \mid \mathcal{M}_2)} \\
&= \frac{1/27}{1/27 + 1/27} \\
&= 1/2 \\
P(L \mid W, \mathcal{M}_3) &= \frac{P(L, W \mid \mathcal{M}_3)}{P(W \mid \mathcal{M}_3)} \\
&= \frac{1/6}{1/6 + 1/18} \\
&= 3/4
\end{aligned}$$

At this point, substitute the second term on the right hand side, the just recently completed computational chore of finding the posterior probability for each model,

$$\begin{aligned} P(L \,|\, W) &= (1/2 \times 3/7) + (1/2 \times 1/7) + (3/4 \times 3/7) \\ &= 17/28 \end{aligned}$$

Compare this conditional probability of life $P(L \,|\, W) = 17/28 \approx 61\%$ with the marginal probability of life from the joint probability table of $P(L) = 17/54 \approx 31\%$. Evidently, if water were to be discovered, an IP's degree of belief in the truth of the statement that "Life exists on Mars." is raised dramatically.

Of course, such conclusions always have to be blanketed in caveats, foremost being that the argument was conducted in the space of only three models, all of which implemented $L \to W$ to some degree or another. On the other hand, these kind of numerical exercises are very helpful in providing quantitative insight into just how much a degree of belief will change when the rules of probability theory are strictly followed. Perhaps more importantly, the results from such exercises serve to stimulate further debate focused on what is now revealed to be the missing links supporting an improved argument about the presence of life on Mars if water were to be discovered.

Supplemental Exercise 7.3.4: How would the IP more typically approach the construction of the joint probability tables for the Life on Mars scenario?

Solution to Supplemental Exercise 7.3.4:

Instead of explicitly folding in the model statements in order to construct just one $n = 12$ joint probability table, the IP would most likely construct several $n = 2 \times 2 = 4$ joint probability tables, each explicitly conditioned on $\mathcal{M}_k$. Then, for example, the one $n = 4$ table under model $\mathcal{M}_3$ would have the numerical assignments of 1/2, 1/6, 0, and 1/3 replacing the assignments seen in cells 9, 10, 11, and 12. The assignments for each $n = 4$ table would have to sum to 1.

Of course, the computations as done above are altered only to the extent that, for example, focusing on the $n = 4$ table constructed under model $\mathcal{M}_1$, where the assignments in cells 1, 2, and 4 are now all 1/3,

$$P(W, \mathcal{M}_1) = P(W \,|\, \mathcal{M}_1) \times P(\mathcal{M}_1) = (1/3 + 1/3) \times 1/3 = 2/9$$

The denominator in Bayes's Theorem then changes to,

$$P(W) = \sum_{k=1}^{3} P(W \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) = \frac{2}{9} + \frac{2}{27} + \frac{2}{9} = \frac{6}{27} + \frac{2}{27} + \frac{6}{27} = \frac{14}{27}$$

but the updated probability for the models conditioned on the assumed presence of water must remain the same at,

$$P(\mathcal{M}_1 \,|\, W) = \frac{P(\mathcal{M}_1, W)}{P(W)} = \frac{6/27}{14/27} = \frac{3}{7}$$

$$P(\mathcal{M}_2 \,|\, W) = \frac{P(\mathcal{M}_2, W)}{P(W)} = \frac{2/27}{14/27} = \frac{1}{7}$$

$$P(\mathcal{M}_3 \,|\, W) = \frac{P(\mathcal{M}_3, W)}{P(W)} = \frac{6/27}{14/27} = \frac{3}{7}$$

The three conditional probabilities $P(L \,|\, W, \mathcal{M}_k)$ also remain the same, as for example under model $\mathcal{M}_1$, $P(L \,|\, W, \mathcal{M}_1) = \frac{1/3}{1/3+1/3} = 1/2$. Our ultimate objective, the final inference concerning the presence of life given the presence of water, has not been affected either with $P(L \,|\, W) = 17/28$.

7.4 Where are the Data?

I would like to finish up with a conceptual point. In these exercises generalizing logic with probability, you may have noticed that no data were required. Or, equivalently, we didn't write out any probability expressions like $P(L \,|\, W, \mathcal{D})$. The established truth of the conclusion W and the acceptance of the logic function as a model $L \to W$ were all that were necessary for the inference.

Neither did probability expressions involving results over repeated trials appear. The necessity of somehow resolving this issue for the probability over repeated trials in the Bernoulli Urn Scenario led to the extended discussion in the previous Chapter.

My point being that there is a certain pristine simplicity when probability theory is employed to generalize logic. Not having to deal with any data and not having to engage in a rather involved analysis as Jaynes was forced to did in sampling without replacement leads to a very straightforward application of Bayes's Theorem. Once the presence of water on Mars was confirmed, the rest of the inferential scenario was child's play.

As we have seen, the placement of 0s into a joint probability table by some logic function can be very helpful. On the other hand, we realize that logic functions can also be quite restrictive in this regard.

In addition, as we had occasion to witness in our solution to the logic puzzle, placement of 0s as assigned numerical values to probabilities might lead to confusion further on down the road. This occurs if we haven't been paying sufficient attention in setting up Bayes's Theorem by conditioning on the truth of a statement, together with a model, whne the assertion that both are true happens to be contradictory.

A satisfying conclusion to this introduction of generalizing logic with probability is that there is an obvious fix to problematic 0s. For example, if the IP were to replace the problematic 0s in cells 5 and 7 of the second logic puzzle's joint probability table with some very small number under another model $\mathcal{M}_\star$, then,

$$\begin{aligned} P(R \,|\, \overline{B}, P, \mathcal{M}_\star) &= \frac{P(\overline{B}PR \,|\, \mathcal{M}_\star)}{P(\overline{B}PR \,|\, \mathcal{M}_\star) + P(\overline{B}P\overline{R} \,|\, \mathcal{M}_\star)} \\ &= \frac{\text{Cell 5}}{\text{Cell 5} + \text{Cell 7}} \\ &= \frac{0.001}{0.001 + 0.001} \\ &= 1/2 \end{aligned}$$

An IP's degree of belief in the truth that taxes will be raised given that the budget will not be cut and prices will remain stable is no longer undecidable, but hovers around that mid–point for a binary statement indicating it could go either way.

Of course, if small probabilities were to be assigned to cells 5 and 7, then the other cells with positive probability would have to be adjusted downwards to take account of this fact.

Chapter 8

Deterministic Cellular Automata

8.1 Enforcing Determinism

Wolfram's cellular automata are, by definition, already deterministic. Remember that these CA are stand–ins for an ontological description of how the world operates at the physical level.

I decided to add the redundant adjective to the title of Volume I's Chapter Eight because of the desire to pinpoint the differences that would transform a *deterministic* CA into a *probabilistic* CA. This was the preliminary ground work necessary to Chapter Nine's discussion of probabilistic CA. My version differed conceptually from Wolfram's interpretation, and it is important to understand why this is so.

A little bit of analysis based on Bayes's Theorem quickly pointed the way to the general template for the probability of a white or black cell at the next time step to be 0 or 1. In the notation adopted for the elementary cellular automata, the probability for the color of a cell at the next time step was conditioned on the colors of the three neighboring cells at the previous time step. This probability also depended on assuming one of the ECA true. If the 16 cell joint probability tables were suitably constructed such that the 0s dictated by the DNF of the logic function used as a model were balanced off by 0s elsewhere, then determinism could be enforced through probability.

As a review, this enforcement of determinism happens through the mechanism of Bayes's Theorem.

$$P(B_{N+1} \,|\, A_N, B_N, C_N, \mathcal{M}_k) =$$

$$\frac{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_k)}{P(B_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_k) + P(\overline{B}_{N+1}, A_N, B_N, C_N \,|\, \mathcal{M}_k)}$$

If the numerical assignment for the probability of the joint statement in the numerator were 0, then the probability for the color of the updated cell to be black is 0, or equivalently, the probability is 1 that it is white. The placement of a zero here in the numerator would arise from the analysis of the DNF for the three variable logic function serving as the model. If, instead, the numerical assignment for the second term in the denominator were 0, then the probability for the color of the updated cell to be black is 1, or equivalently, the probability is 0 that it is white. The placement of a zero in the second term of the denominator would be dictated through a symmetry with the original DNF.

Supplemental Exercise 8.1.1: Give an alternative, and, perhaps better answer to Exercise 8.6.1 of Volume I.

Solution to Supplemental Exercise 8.1.1:

The above mentioned exercise was originally intended to introduce an important aspect of probability's formal manipulation rules. This generality covered the case where models could also be included as statements within probability expressions. Therefore, I returned to the generic A, B, C format with the idea that statement C would play the role of any model.

For the purposes of the exercise, C was the statement that the `NAND` logic function was the inspiration for the numerical assignments in some joint probability table. As a consequence, model $\overline{C}$ must represent the mirror image of model C where the coefficients in the DNF of `NAND` are reversed. This turns out to be the logic function `AND`.

The IP's goal is to update its degree of belief about the truth of the model represented by C when conditioned on the fact that A is TRUE and B is TRUE. The formal manipulation rule is Bayes's Theorem,

$$P(C \mid A, B) = \frac{P(CAB)}{P(AB)}$$

Expand the denominator to yield,

$$P(C \mid A, B) = \frac{P(CAB)}{P(CAB) + P(\overline{C}AB)}$$

Construct an acceptable eight cell joint probability table for this situation of three binary variables where the numerical assignments of 0 are dictated by the coefficient $f(T, T) = F$ in the DNF of `NAND`, and the mirror image coefficients for `AND`, $f(F, T) = f(T, F) = f(F, F) = F$. See Figure 8.1 at the top of the next page.

After this much has been accomplished, it is clear that,

$$P(C \mid A, B) = \frac{P(ABC)}{P(ABC) + P(AB\overline{C})} = \frac{0}{0 + 1/2} = 0$$

	C				$\overline{C}$			
	A	$\overline{A}$			A	$\overline{A}$		
B	0 (1)	1/6 (2)	1/6	B	1/2 (5)	0 (6)	1/2	2/3
$\overline{B}$	1/6 (3)	1/6 (4)	1/3	$\overline{B}$	0 (7)	0 (8)	0	1/3
	1/6	1/3	1/2		1/2 2/3	0 1/3	1/2	1

Figure 8.1: *A joint probability table for the* `NAND` *and* `AND` *logic functions.*

and,

$$P(\overline{C} \mid A, B) = \frac{P(AB\overline{C})}{P(AB\overline{C}) + P(ABC)} = \frac{1/2}{1/2 + 0} = 1$$

The universe of discourse only included two models $\mathcal{M}_1$ and $\mathcal{M}_2$. The rules of probability ruled out model $\mathcal{M}_1$ as impossible, leaving model $\mathcal{M}_2$ as the sole survivor. If it should happen that both A and B could be FALSE under this surviving model, then the IP realizes that it started out with an impoverished set of models.

Supplemental Exercise 8.1.2: Write out the requisite *Mathematica* code mentioned in Exercise 8.6.4 of Volume I proving the logical equivalency between the shorter and longer DNF for the `NAND` function.

Solution to Supplemental Exercise 8.1.2:

The fully expanded DNF for `NAND` is,

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

First evaluate, $\overline{A \wedge B} \leftrightarrow (A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$

```
TautologyQ[Equivalent[Not[And[a, b]],
     Or[And[a, Not[b]], And[Not[a], b], And[Not[a], Not[b]]]]]
```

followed by, $\overline{A} \vee \overline{B} \leftrightarrow (A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B})$

```
TautologyQ[Equivalent[Or[Not[a], Not[b]],
     Or[And[a, Not[b]], And[Not[a], b], And[Not[a], Not[b]]]]]
```

Since both evaluations return **`True`**, we can use the shorter DNF for `NAND` of either $\overline{A} \vee \overline{B}$ or $\overline{A \wedge B}$.

Supplemental Exercise 8.1.3: Relying upon the new joint probability table of Supplemental Exercise 8.1.1, calculate $P(\overline{B}\,|\,\overline{A})$ just as it was done in Exercise 8.6.6 of Volume I.

Solution to Supplemental Exercise 8.1.3:

Because it is such a very important concept in Bayesian statistics, let us review the formula implementing averaging with respect to a posterior probability for models. From the fundamental principles borrowed from Boolean Algebra, we start with,

$$P(\overline{A}, \overline{B}, \mathcal{M}_k) = P(\overline{B}, \mathcal{M}_k, \overline{A})$$

Invoking the **Sum Rule**,

$$P(\overline{A}, \overline{B}) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}, \mathcal{M}_k, \overline{A})$$

At the next step, invoke the **Product Rule**,

$$P(\overline{A}, \overline{B}) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}\,|\,\mathcal{M}_k, \overline{A})\, P(\mathcal{M}_k\,|\,\overline{A})\, P(\overline{A})$$

Now, divide both sides by $P(\overline{A})$,

$$\frac{P(\overline{A}, \overline{B})}{P(\overline{A})} = \sum_{k=1}^{\mathcal{M}} P(\overline{B}\,|\,\mathcal{M}_k, \overline{A})\, P(\mathcal{M}_k\,|\,\overline{A})$$

The left hand side is the very definition of Bayes's Theorem, so we have,

$$P(\overline{B}\,|\,\overline{A}) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}\,|\,\mathcal{M}_k, \overline{A})\, P(\mathcal{M}_k\,|\,\overline{A})$$

Simply for aesthetic reasons, choose to display this last formula by switching $\mathcal{M}_k$ and $\overline{A}$ around in the first term on the right hand side,

$$P(\overline{B}\,|\,\overline{A}) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}\,|\,\overline{A}, \mathcal{M}_k)\, P(\mathcal{M}_k\,|\,\overline{A})$$

It is obvious at this point that the degree of belief that statement B is FALSE when conditioned on knowing that A is FALSE is quite literally an average of all the degrees of belief for this situation. This averaging, however, is performed with respect to the probability distribution of the models after their reorientation knowing that A is FALSE. However, note that $\overline{A}$ was not deleted from the first term on the right hand side. This not a question of invoking independence since the state space was defined to include all joint statements about A and B.

Now for the calculation of $P(\overline{B}\,|\,\overline{A})$ based on the new joint probability table in Figure 8.1. Each of the two terms on the right hand side are themselves calculated by Bayes's Theorem,

$$P(\overline{B}\,|\,\overline{A},\mathcal{M}_k) = \frac{P(\overline{A},\overline{B}\,|\,\mathcal{M}_k)}{P(\overline{A}\,|\,\mathcal{M}_k)}$$

$$P(\mathcal{M}_k\,|\,\overline{A}) = \frac{P(\overline{A}\,|\,\mathcal{M}_k)\,P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\overline{A}\,|\,\mathcal{M}_k)\,P(\mathcal{M}_k)}$$

Here we see the utility of my insistence on retaining the terms in Bayes's Theorem as *joint probabilities* conditioned on the information in some given model. We can read off the numerical assignments to $P(\overline{A},\overline{B}\,|\,\mathcal{M}_1)$ in cell 4, and $P(\overline{A},\overline{B}\,|\,\mathcal{M}_2)$ in cell 8 directly. Likewise, $P(\overline{A}\,|\,\mathcal{M}_1)$ can be read directly as the sum over cells 2 and 4, and $P(\overline{A}\,|\,\mathcal{M}_2)$ as the sum over cells 6 and 8.

Furthermore, since we can easily discern from the joint probability table that $P(\overline{A},\overline{B}\,|\,\mathcal{M}_2) = 0$, we need only calculate,

$$P(\overline{B}\,|\,\overline{A}) = \frac{P(\overline{A},\overline{B}\,|\,\mathcal{M}_1)}{P(\overline{A}\,|\,\mathcal{M}_1)} \times \frac{P(\overline{A}\,|\,\mathcal{M}_1)\,P(\mathcal{M}_1)}{P(\overline{A}\,|\,\mathcal{M}_1)\,P(\mathcal{M}_1) + P(\overline{A}\,|\,\mathcal{M}_2)\,P(\mathcal{M}_2)}$$

from the information in the first model.

$$\begin{aligned} P(\overline{B}\,|\,\overline{A}) &= \frac{1/6}{1/3} \times \frac{(1/3 \times 1/2)}{(1/3 \times 1/2) + (0 \times 1/2)} \\ &= \frac{1}{2} \end{aligned}$$

This is the same degree of belief that B is FALSE as in Exercise 8.6.6 when different numerical assignments were made under different models. Both sets of models were, nonetheless, inspired by the `NAND` and `AND` logic functions in their placement of the 0s in the joint probability table.

After the fact, it is relatively easy to talk ourselves into a rationale justifying such a probability. We began with only two models, but one of them was ruled out immediately given that we are conditioning on the truth of $\overline{A}$. Within the one surviving model, $\overline{A}$ can indeed happen. Given the truth that A is FALSE, either B or $\overline{B}$ have equal probabilities under the information from the one remaining model. Therefore, we would want to lend equal credibility to both propositions. Hence, a degree of belief $P(\overline{B}\,|\,\overline{A}) = P(B\,|\,\overline{A}) = 1/2$.

Supplemental Exercise 8.1.4: Give an example of an inference that is certain because it is based on one of Wolfram's 256 elementary cellular automata.

Solution to Supplemental Exercise 8.1.4:

This supplemental exercise is very similar to Volume I's Exercise 8.6.10. There, I cataloged all of the three variable logic functions $f_\star(A, B, C)$ with respect to the number of 0s in their joint probability tables. We found that there were a total of,

$$\binom{8}{4} = 70$$

such logic functions with four zeroes scattered somewhere around an eight cell joint probability table.

Wolfram's 256 elementary cellular automata run according to some rule that is equivalent to one of these three variable logic functions. Any of these 256 three variable logic functions, $f_\star(A, B, C)$, except for the FALSE logic function, might also serve as a model for making numerical assignments to the joint statements for some defined state space.

Suppose that we actually begin the enterprise from the standpoint of the joint probability table. As usual, we will have a model $\mathcal{M}_k$ whose information will direct the numerical assignments to all eight cells indexing the probability for some joint statement. In symbols, then, we start with the familiar expressions of $P(A, B, C \,|\, \mathcal{M}_k)$ in cell 1 through $P(\overline{A}, \overline{B}, \overline{C} \,|\, \mathcal{M}_k)$ in cell 8.

We said that we would like to investigate one of the seventy logic functions that would place four zeroes somewhere among the available eight cells. Let's choose one where the zeroes are located in cells 2, 3, 5, and 8. Examine the resulting joint probability table in Figure 8.2 at the top of the next page.

Quite obviously, what we have here is, in fact, a 16 cell joint probability table with an additional variable B_{N+1}. The first eight cells have the four zeroes located according to our specification given above. The second eight cells also have four zeroes located such that the probability calculations for the color of the cell at the next time step $B_{N+1} = T$ or F result in 0 or 1. Wherever we find the coefficients in the expansion of the logic function that are equal to T for the first eight cells, we reverse this and make the coefficients equal to F for the second eight cells.

Also, pay careful attention to the labeling of the rows and columns corresponding to statements B and C. This is because we want to follow *Mathematica*'s ordering convention from the list produced by **`Tuples[{True, False}, 3]`** which contains the eight elements in this order,

$$TTT, TTF, TFT, TFF, FTT, FTF, FFT, FFF$$

Thus, moving over to say the third element TFT, we would want this to correspond to cell 3 indexing originally $P(A\overline{B}C \,|\, \mathcal{M}_k)$, but now $P(B_{N+1}, A_N, C_N, \overline{B}_N \,|\, \mathcal{M}_k)$.

B_{N+1}									
	A_N			$\overline{A}_N$					
	C_N	$\overline{C}_N$			C_N	$\overline{C}_N$			
B_N	1/8 (1)	0 (2)	1/8	B_N	0 (5)	1/8 (6)	1/8	1/4	
$\overline{B}_N$	0 (3)	1/8 (4)	1/8	$\overline{B}_N$	1/8 (7)	0 (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	0 (9)	1/8 (10)	1/8	B_N	1/8 (13)	0 (14)	1/8	1/4	1/2
$\overline{B}_N$	1/8 (11)	0 (12)	1/8	$\overline{B}_N$	0 (15)	1/8 (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

Figure 8.2: *A joint probability table based on Rule 150 and its complement to enforce inferences about the color of cells that have probabilities of 0 or 1.*

Since we have specified the cells where we want the four zeroes, we can write down the binary number associated with Wolfram's rule number for the analogous ECA. Table 8.1 shows that this particular choice of zeroes for a joint probability table corresponds to Wolfram's Rule 150.

What does Wolfram give as the logic function behind Rule 150? It is expressed in his notation $\veebar$ for the **`Xor[ ]`** function as,

$$\text{Rule 150} \Rightarrow p \veebar q \veebar r$$

I always prefer the fully expanded DNF version for any logic function such as $f_{150}(A, B, C)$. One reason is simply that I always want to force the canonical form into symbolic expressions involving only $\wedge$ and $\vee$. It is easy to read off where the function coefficients are equal to T in Table 8.1 together with their matching building block orthogonal functions. The DNF for $f_{150}(A, B, C)$ must look like this,

$$\text{Rule 150} \Rightarrow ABC \vee A\overline{B}\,\overline{C} \vee \overline{A}B\overline{C} \vee \overline{A}\,\overline{B}C$$

Table 8.1: *Translating the placement of 0s in the eight cell joint probability table into Wolfram's rule number for an ECA.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
T	F	F	T	F	T	T	F
1	0	0	1	0	1	1	0
2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
128	0	0	16	0	4	2	0
Rule number 150							

Pause here to have *Mathematica* verify these identities. Examine the output from the following code, executed in this order, to see the series of steps bringing us to the identity between Wolfram's logic expression and my fully expanded DNF.

```
1.  f = BooleanFunction[150, 3]
2.  BooleanConvert[f[a, b, c]] // FullForm
3.  BooleanTable[f[a, b, c]]
4.  TautologyQ[Equivalent[f[a, b, c], Xor[a, b, c]]]
```

With all of this preliminary groundwork completed, it is straightforward to make inferences concerning the truth of statements about the color of a cell at the next time step. These inferences will always output probabilities of 0 or 1. Thus, an IP will exhibit no doubt in its inferences about whether the cell in question is colored white or black.

For example, if the three neighboring cells at the previous time step were all black, then the cell under consideration will also be colored black with probability 1.

$$\begin{aligned} P(B_{N+1} \,|\, A_N, B_N, C_N, \text{Rule 150}) &= \frac{P(B_{N+1}, A_N, B_N, C_N)}{P(B_{N+1}, A_N, B_N, C_N) + P(\overline{B}_{N+1}, A_N, B_N, C_N)} \\ &= \frac{\text{Cell 1}}{\text{Cell 1} + \text{Cell 9}} \\ &= \frac{1/8}{1/8 + 0} \\ &= 1 \end{aligned}$$

On the other hand, if the cell immediately above were white and its two neighbors were black, then the cell under consideration is colored black with probability 0, or colored white with probability 1. For less clutter, the conditioning on the model is not shown on the right hand side of the formulas.

$$
\begin{aligned}
P(B_{N+1} \mid A_N, \overline{B}_N, C_N, \text{Rule 150}) &= \frac{P(B_{N+1}, A_N, \overline{B}_N, C_N)}{P(B_{N+1}, A_N, \overline{B}_N, C_N) + P(\overline{B}_{N+1}, A_N, \overline{B}_N, C_N)} \\
&= \frac{\text{Cell 3}}{\text{Cell 3} + \text{Cell 11}} \\
&= \frac{0}{0 + 1/8} \\
&= 0
\end{aligned}
$$

8.2 More Complex Deterministic CA

In my examination of deterministic CA in Chapter Eight of Volume I, I never strayed from Wolfram's set of 256 elementary CA. But it is certainly feasible as an application of Bayes's Theorem's ability to reproduce probabilities of 0 or 1 for the next occurrence of a black or white cell to look at more complicated CA.

The topic was broached at the end of Chapter Three through the cursory mention of how to go about generalizing CA. The 256 elementary CA all updated the current cell's color by looking at the color of that cell at the previous time step as well as its immediate nearest neighbors, the one cell on the left and the one cell on the right. These causal variables were given the notation of A_N, B_N, and C_N with the unknown variable to be predicted labeled as B_{N+1}.

Chapter Three introduced a CA that took as causal variables the colors of the *two* nearest neighbors to the cell being updated. So the notation changes to finding a probability $P(C_{N+1} \mid A_N, B_N, C_N, D_N, E_N, \mathcal{M}_k)$.

The following discussion in this introductory section repeats many of the ideas that first appeared back in Supplemental Exercise 3.2.6. But as I have emphasized in my *Apologia*, I never shy away from going over plowed ground. I have discovered that there is always some pedagogical merit in repetition.

The set of 256 elementary CA mapped directly over to logic functions of three propositions. Wolfram's formula to find the total number of CA finds that,

$$k^{k^{(2r+1)}} = 2^{2^3} = 256$$

where $k = 2$ stands for the number of colors, and $r = 1$ stands for the *range* of neighbors. This formula also tells us the total number of logic functions where now $k = 2$ describes the carrier set consisting of only TRUE and FALSE and $2r + 1$ describes the number of arguments to the logic functions.

Just as in the example used in Chapter Three, for our current example of a more complicated CA, keep the number of colors at two, black and white, or, the

elements in the carrier set at two, TRUE and FALSE. However, let's generalize to $r = 2$ nearest neighbors or $2r + 1 = 5$ arguments to any function so that a logic function would be written as $f_\star(A, B, C, D, E)$. There is a vast increase in the possible number of logic functions to over four billion,

$$k^{k^{(2r+1)}} = 2^{2^5} = 4,294,967,296$$

Remember that our focus is on the computation of the probability of the color of the updated cell in this more complicated CA through Bayes's Theorem,

$$P(C_{N+1} \,|\, A_N, B_N, C_N, D_N, E_N, \mathcal{M}_k)$$

Thus, given the five causal variables and the one variable to be predicted, a joint probability table with $n = 64$ cells will have to be constructed. Each one of these 64 cells will contain a numerical assignment to a joint statement generically rendered as $(C_{N+1}, A_N, B_N, C_N, D_N, E_N)$. The numerical assignment will ensue from the information resident in some model $\mathcal{M}_k$, or, equivalently, from assuming the truth of one of the four billion logic functions and its complement.

Exploring this type of inferential problem involving complicated CA is made much easier with the availability of *Mathematica* functions like,

```
IntegerDigits[ ], BooleanFunction[ ], BooleanConvert[ ]
```

and so on. And when we want to actually look at any complicated CA in all of its evolving glory we, of course, will rely on,

```
ArrayPlot[CellularAutomaton[··· ]]
```

Supplemental Exercise 8.2.1: Review the specification of the rule number that will be inserted as an argument into `CellularAutomaton[ ]` for our example of a complicated CA.

Solution to Supplemental Exercise 8.2.1:

Since there are a total of 4,294,967,296 possible rule numbers, or logic functions, any rule number according to Wolfram's specifications must be in the range from 0 to 4,294,967,295. Each rule number is decomposed into base 2 notation. Thus, any rule number will consist of $2^5 = 32$ digits of either 0 or 1. As anchor points, the list of digits for rule number 4,294,967,295 is found through,

```
IntegerDigits[4 294 967 295, 2, 32]
```

The list consists of 32 1s as digits.

At the other extreme, the list of digits for rule number 0 is found through,

```
IntegerDigits[0, 2, 32]
```

The list consists of 32 0s as digits.

Somewhere in the middle is the rule number used as the example in Chapter Three of Volume I.

```
IntegerDigits[2 147 483 649, 2, 32]
```

The list consists of 32 binary digits with the first and last digits 1 and the remaining thirty digits all 0.

Recall that these binary digits match up with a functional assignment of TRUE or FALSE. We wanted this logic function of five arguments $f_\star(A, B, C, D, E)$ to have a functional assignment of T at $f_\star(A, B, C, D, E)$, a functional assignment of T at $f_\star(\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E})$, and functional assignments of F everywhere else.

Supplemental Exercise 8.2.2: Select some rule number to act as the model for the numerical assignments to a 64 cell joint probability table.

Solution to Supplemental Exercise 8.2.2:

Arbitrarily choose some number in the range from 0 to 4,294,967,295. Back in Supplemental Exercise 3.2.6, I picked the number 3,916,273,040. This is as good a rule number as any, so we will continue to use it.

Let's examine the resulting list of 32 binary digits so that we can identify where the functional assignments of F and T take place.

```
IntegerDigits[3 916 273 040, 2, 32]
```

returns the explicit list of 32 "1"s and "0"s,

```
{1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0,
 1, 1, 0, 0, 1, 0, 0, 0, 0}
```

With this result, we know how many Ts and Fs we have, but we don't know where to place them in the joint probability table. For that we need to examine the output from `Tuples[{T, F}, 5]`.

This will produce a list of lists of 32 elements with each of the 32 elements itself a list of five elements all T or F. The first list of five elements all T or F is matched up with the first cell of the joint probability table, the second list of five elements all T or F is matched up with the second cell of the joint probability table, and so on until all 32 lists are matched up with the first 32 cells of the 64 cell joint

probability table. There is a second type of matching up going on concomitantly with the list of 1s and 0s from **IntegerDigits[ ]** determining whether a 0 or legitimate probability value is assigned to that cell.

So, for example, since the outcome from **Tuples[ ]** looks like this,

```
{{T, T, T, T, T}, {T, T, T, T, F}, ···,
                                   {F, F, F, F, T}, {F, F, F, F, F}}
```

the first cell indexes the probability for the joint statement $(ABCDE)$, the second cell indexes the probability for the joint statement $(ABCD\overline{E})$, $\cdots$, the next to last cell indexes the probability for the joint statement $(\overline{A}\,\overline{B}\,\overline{C}\,\overline{D}E)$, while the last cell indexes the probability for the joint statement $(\overline{A}\,\overline{B}\,\overline{C}\,\overline{D}\,\overline{E})$. The subscript N which should be attached to each statement is omitted for clarity.

The second match up shows that a legitimate probability will appear in the first three cells, a zero in the fourth cell, and so on, until the last cell which also is assigned a zero. For the bottom half of the table, the last 32 cells indexing C_{N+1}, make the complementary assignment. Taking all of this into account, we can construct our 64 cell joint probability table, sketched in Figure 8.3 at the top of the next page.

Any probability for the color of an updated cell in this more complicated CA can now be calculated as black or white with certainty. This is the hallmark of a deterministic CA, or, more germane to our goal, another example of how probability theory generalizes logic.

Suppose the current cell to be updated is C_{N+1}, and its immediate predecessor is black, while the two neighbors to its right are black and white, and the two neighbors to the left are also black and white. What is the IP's degree of belief in the truth of the statement that C_{N+1} will be colored black? Bayes's Theorem computes,

$$P(C_{N+1} \,|\, A_N, \overline{B}_N, C_N, \overline{D}_N, E_N, \mathcal{M}_k) = 1$$

so the IP is certain that the cell to be updated will be black.

Examine the joint probability table containing the numerical assignments under model $\mathcal{M}_k$ in Figure 8.3 to locate the two relevant joint probabilities needed for an application of Bayes's Theorem. Cell 11 indexes,

$$P(C_{N+1}, A_N, \overline{B}_N, C_N, \overline{D}_N, E_N \,|\, \mathcal{M}_k)$$

while cell 43 indexes the complementary assignment for,

$$P(\overline{C}_{N+1}, A_N, \overline{B}_N, C_N, \overline{D}_N, E_N \,|\, \mathcal{M}_k)$$

Bayes's Theorem then tells us that,

$$P(C_{N+1} \,|\, A_N, \overline{B}_N, C_N, \overline{D}_N, E_N, \mathcal{M}_k) = \frac{\text{cell 11}}{\text{cell 11} + \text{cell 43}} = \frac{q}{q+0} = 1$$

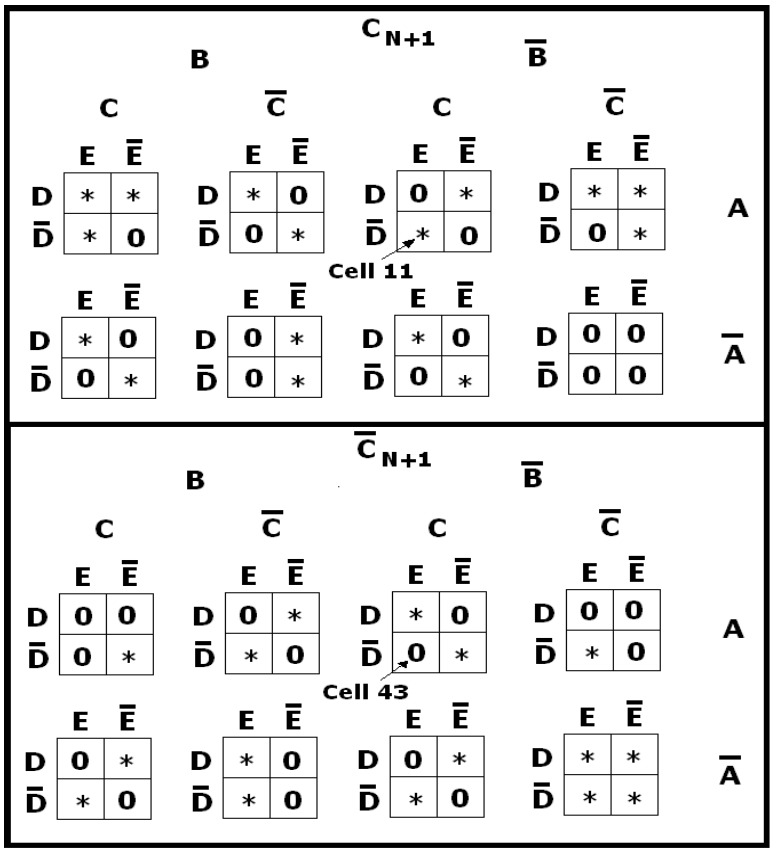

Figure 8.3: *A joint probability table based on a rule number for a more complicated CA involving* two *nearest neighbors. A model based on this rule and its complement allows inferences about the color of cells that have probabilities of 0 or 1.*

Supplemental Exercise 8.2.3: Confirm this certain output of a black cell from an application of Bayes's Theorem by viewing the full set of 32 transformations under Rule Number 3 916 273 040.

Solution to Supplemental Exercise 8.2.3:

Through the simple expedient of,

```
RulePlot[CellularAutomaton[{3 916 273 040, 2, 2}]]
```

we can confirm that **BWBWB** produces a black cell as shown in Figure 8.4 on the next page.

The top of each block reproduces the output we saw from `Tuples[ ]`. For example, the fourth block over, the last one in the first row, shows **BBBWW**. The corresponding output from `Tuples [ ]` was `{T, T, T, F, F}`. The color of

the updated cell at the bottom of each block matches **`IntegerDigits[ ]`** which has a 0 at the fourth position over from the left. Notice also that our example transformation via Bayes's Theorem occurs as the 11th one listed, marked with an $\star$, matching its placement within the joint probability table.

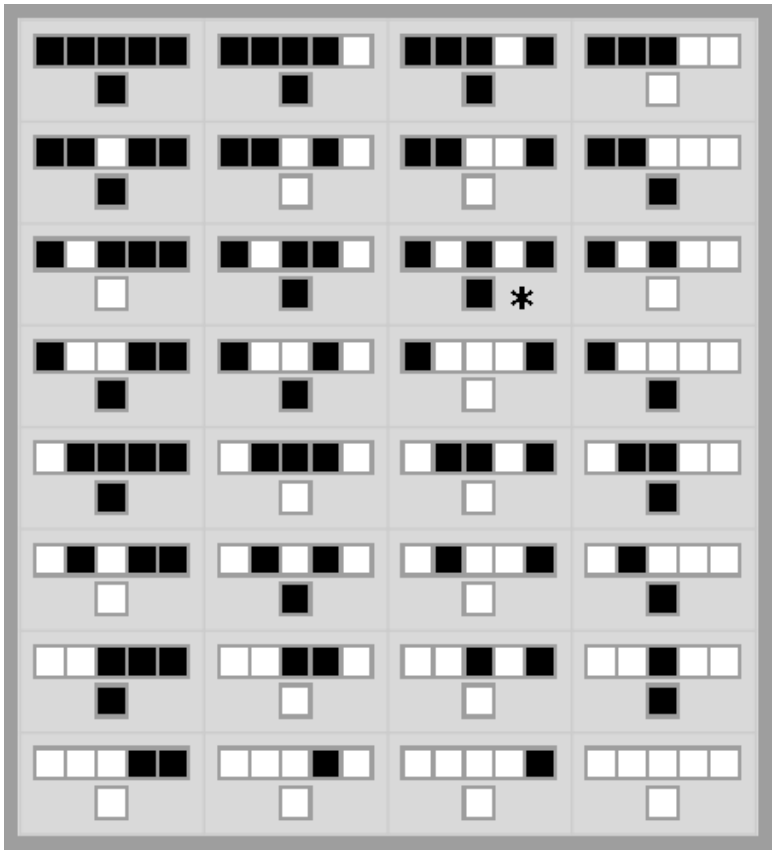

Figure 8.4: **`RulePlot[ ]`** *diagram for the complicated CA.*

Supplemental Exercise 8.2.4: Show this more complicated deterministic CA evolving over 20 time steps with some specific starting conditions.

Solution to Supplemental Exercise 8.2.4:

Once again, Wolfram has provided all the tools in *Mathematica* for examining all sorts of complicated CA. To see our example evolving over 20 steps with an initial arbitrary starting condition of black and white cells,

```
ArrayPlot[CellularAutomaton[{3 916 273 040, 2, 2},
              {1, 0, 0, 0, 1, 0, 1, 0, 1, 0}, 20], Mesh → True]
```

The first list as an argument to **`CellularAutomaton[ ]`** has three elements, the rule number, the number of colors, and the range. Meanwhile, the second list as an argument contains the colors of the starting configuration of cells at the beginning time step. The final argument specifies the number of steps into the future we want the CA to evolve.

Figure 8.5 below shows the evolution of the CA determined by rule number 3,916,273,040, randomly selected from the total possible 4,294,967,295 rule numbers, as produced by **`ArrayPlot[ ]`**,

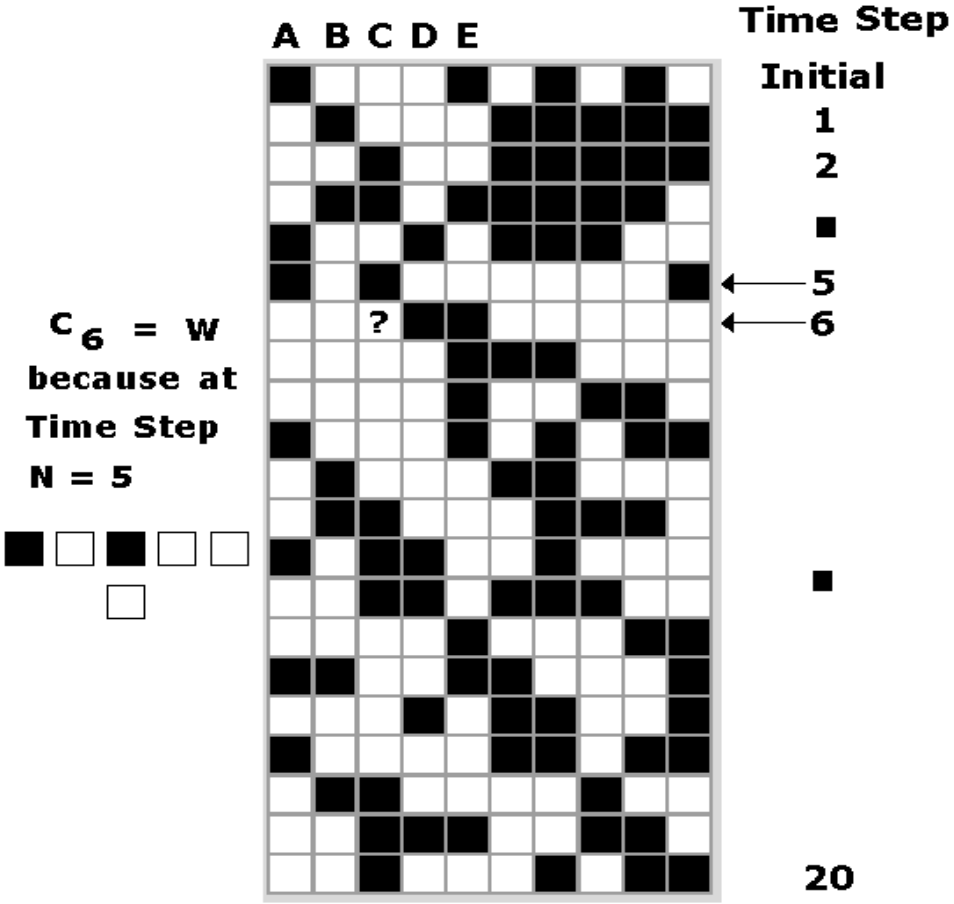

Figure 8.5: *The evolution of a more complicated deterministic CA for 20 steps.*

Chapter 9

Probabilistic Cellular Automata

9.1 Probabilistic CA: An Oxymoron

Just as I characterized the descriptive term *deterministic cellular automata* in the last Chapter as redundant, it then follows that here I must now characterize the descriptive term *probabilistic cellular automata* as an oxymoron. I think that the juxtaposition of the two concepts *physically determined* and *epistemic probability* qualifies as an oxymoron.

Stephen Wolfram is flat–out wrong when he tries to blend in probability as a way to generalize his cellular automata. He mentions that the color of each cell in some CA could be assigned with some fixed independent probability p. So, for example, if two neighboring cells at the previous time step were black and white, then the current cell might be assigned the color black with probability $p = 1/2$. Unfortunately, cells in a CA just don't "have a probability p" as one of their physical properties.

This is why Jaynes was shocked to his core when he realized just how rampant was this entrenched and thoroughly unexamined concept that probabilities reflected physical properties rather than a state of knowledge. The ludicrous idea that a human's state of knowledge as represented by a probability could influence the physics of the real world was labeled by him as the *Mind Projection Fallacy*.

The *Mind Projection Fallacy* views a probability p as just another physical property of some object; in this case, determining whether a cell is colored white or black. Unfortunately, a probability **CAN NEVER** be conceptualized as some physical property possessed by an object.

The fact that Quantum Mechanics does seem to fold in probability as a physical property means that some serious conceptual reorientation is in order. I suggest that substituting the phrase "physical propensity," but retaining the attendant mathematics of complex numbers to replace the phrase "probability amplitudes."

A probability **CAN ONLY** be conceptualized as representing a *degree of belief* as held by an IP in the truth of a statement that "this object possesses this physical property." An IP's degree of belief in such statements is always conditioned on the information resident in some model $\mathcal{M}_k$.

Or, if you prefer, turn this around and say that entropy quantifies the amount of *missing information* for an IP's state of knowledge. This is the basis for the MEP's assignment of numerical values to probabilities of joint statements conditioned on the information in some model as examined in detail in Volume II.

In addition, whether a cell in a CA is allowed to take on a continuous shades of gray as opposed to just black or white has nothing whatsoever to do with probability. Whatever colors a cell might possess is an ontological property. Since probability is an epistemological concept, that is, a state of knowledge as held by an IP, a probability cannot influence by one iota the ontological status of a cell's color.

In section 9.4 of Volume I, I tried to emphasize this fact by pointing out that the carrier set $\mathbf{B}$ of a Boolean Algebra could have elements other than the required elements of T or F that correspond to the dichotomous black and white color of each CA's cell. Add to $\mathbf{B}$ elements a and a' for two shades of gray intermediate between black and white. Continue to add elements b and b' to the carrier set to implement further shades of gray. Perhaps you are a person inclined to like fifty shades of gray.

The resulting "big Boolean Algebra" remains an unadulterated deductive system just like a CA that had added shades of gray as further physical properties of a cell. Nowhere does the concept of probability make an appearance through this ruse of enlarging the carrier set. Nor does probability have to enter into the notion of a deterministic CA just because the set of colors has been enlarged.

In Classical Logic, propositions must be either TRUE or FALSE. Furthermore, their adjudications through some logic function must also fall out as TRUE or FALSE. Probability generalizes logic in that, although the statements still must be either TRUE or FALSE, their adjudication can be satisfied by a probability falling between 0 and 1.

Likewise, for a CA, the cell must truly be colored white, or black, or some shade of gray, or some other color. Any other ontological property of the cell might be defined, and it too must either be TRUE or FALSE for every cell. Demanding that the properties of the cells approach some continuous measurement has nothing to do with probability.

An IP can use probability epistemologically to assert that its degree of belief in the truth of the statement that the cell, which in ontological reality might be

colored black, is colored black with probability 1/2. This cell is definitely one color or another, possesses one physical property or another, but most certainly does not exist in some *superposition* of properties which must eventually be adjudicated through recourse to a probability as Wolfram says. Nature does not consult a random number generator to decide which physical property it will assume at the next instant.

Where Wolfram and I do agree on the correct invocation of probability for CA centers on an almost throwaway statement buried in his Notes for a *A New Kind of Science* on page 922. Here he states:

> **Probabilistic cellular automata.** As an alternative to 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. Examples of probabilistic cellular automata are shown on page 591; their behavior is typically quite similar to continuous cellular automata.

Examples were given in Chapter Nine of Volume I, section 9.3.2, concerning an IP making inferences about the color of the current cell under consideration when it was maximally uncertain about whether it was Rule 110 or Rule 30 governing the operation of the CA. The last two exercises, Exercises 9.5.15 and 9.5.16, went into all of the detail of employing Bayes's Theorem to correctly think about how probability should be brought into the discussion of deterministic CA. For example, the IP's degree of belief in the statement that the cell was colored white was 1/2.

Once again, however, it is very important to realize that ontologically speaking, in the real physical world the CA is emulating, the cell was actually colored black or white, no doubt about it. Epistemologically speaking, the IP was burdened with missing information about which rule was actually governing the color of the cells. Consequently, it was forced to set up a probability distribution over all the models defined to be in the conversation.

So where continuity and probability do begin to enter the fray is through the fact that the IP is provided with a continuous scale from 0 to 1 in order to fashion expressions like,

$$\begin{aligned} P(\mathcal{M}_k \equiv \text{Rule } 0) &= 1/256 \\ P(\mathcal{M}_k \equiv \text{Rule } 1) &= 1/256 \\ \cdots \qquad &= 1/256 \\ P(\mathcal{M}_k \equiv \text{Rule } 255) &= 1/256 \end{aligned}$$

Notwithstanding his note above, Wolfram does not provide any examples in his book of probabilistic cellular automata, as I have done, that proceed directly from his explanation of an alternative way to introduce probabilities. The examples on

his page 591 are exactly the ones I criticized because the cell color is determined probabilistically (that pesky oxymoronic juxtaposition once again). Each cell has the physical property of possessing a probability p that will determine its color.

But this is conceptually wrong. The only thing that can determine the color of the current cell under consideration is the rule. The color of the cell cannot be determined by an IP's degree of belief in the truth of the statement that the cell is a certain color.

Why do I harp on this point *ad nauseam*? Do you want to place your bets in Laplace and determinism? Or would you rather go down the road with quantum mechanics?

The conventional wisdom leaves no room for debate. You are judged to hold antediluvian notions of the worst kind if you have the temerity to suggest that you have not given up on the scientific enterprise.

Trusting in the scientific enterprise, you stubbornly insist that you are going to continue to search for causes. You choose not to entertain the notion that it is the elves and fairies at the bottom of your garden (otherwise known as probabilities as fundamental physical causes) who are responsible for what you see.

9.2 Correct Involvement of Probability

This section presents a concrete example that tries to illustrate more clearly the difference between Wolfram's characterization of how probability gets involved in CA when contrasted with my opposing viewpoint in the discursive discussion as just given above. The example is based on section 9.3.3 in Volume I.

To orient ourselves, let's focus in on Rule 255 for elementary cellular automata. This is a very special rule because it implements in three variables what the **TRUE** logic function did for two variables. The functional assignment $f_\star(A, B, C)$ for all eight possibilities of variable assignments is T in every case. All eight coefficients for the orthogonal basis functions ABC through $\overline{A}\,\overline{B}\,\overline{C}$ are T. There are no F coefficients for this function.

In terms of an elementary cellular automaton, this translates into all cells turning black at the first step after any initial configuration of cell colors. All eight possible color configurations of three cells at the previous time step correspond to all eight orthogonal basis functions, while the functional assignment of T corresponds to the updated cell always colored black for all eight color configurations. The binary number for this rule would then be,

$$11111111 = 2^7 + 2^6 + \cdots + 2^1 + 2^0 = \text{Rule } 255$$

A picture of a cellular automaton running according to Rule 255 is shown on page 56 of a *A New Kind of Science*. After a starting configuration of all white cells with one black cell, all cells are colored black at the very next step. They must remain black for the eternity that Rule 255 is allowed to run.

Supplemental Exercise 9.2.1: What might be a typical joint probability table to enforce determinism under Rule 255?

Solution to Supplemental Exercise 9.2.1:

The first eight cells of a sixteen cell joint probability contain a numerical assignment of 1/8 for the probability of each joint statement starting with B_{N+1}. The last eight cells of the sixteen cell joint probability contain a numerical assignment of 0 for the probability of each joint statement starting with $\overline{B}_{N+1}$. Then, by Bayes's Theorem,

$$P(B_{N+1} = \text{black} \,|\, \text{ any } A_N, B_N, C_N, \mathcal{M}_k \equiv \text{Rule 255}) = 1$$

and,

$$P(B_{N+1} = \text{white} \,|\, \text{ any } A_N, B_N, C_N, \mathcal{M}_k \equiv \text{Rule 255}) = 0$$

Supplemental Exercise 9.2.2: Marginalize over all models to find the probability for the color of the current cell.

Solution to Supplemental Exercise 9.2.2:

If the IP wants to get rid of the conditioning on a specific model on the left hand side of Bayes's Theorem above, then it makes use of the **Sum Rule** to marginalize over all models,

$$P(B_{N+1} = \text{black} \,|\, \text{ any } A_N, B_N, C_N) = \sum_{k=1}^{\mathcal{M}} P(B_{N+1} \,|\, \text{ any } A_N, B_N, C_N, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

If the IP is considering only one model, Rule 255, then $\mathcal{M} = 1$ and $P(\mathcal{M}_{255}) = 1$. Then, the formal manipulation rules of probability will correctly generalize to cover the deterministic scenario,

$$P(B_{N+1} = \text{black} \,|\, \text{ any } A_N, B_N, C_N) = 1$$

Supplemental Exercise 9.2.3: Have the IP introduce probability in the correct way.

Solution to Supplemental Exercise 9.2.3:

Rather than just erroneously asserting that a cell possesses this physical property labeled a probability with, say, a value of $p = 1/2$ that determines its color in a random manner, the rigorous argument demands that probability must enter into any solution in exactly the same consistent manner for every inferencing problem.

For example, suppose that the IP sets up the joint probability table with a numerical assignment of 1/16 to all sixteen cells as in section 9.3.3 of Volume I. Then, a consistent application of the rules of probability reveal that,

$$P(B_{N+1} = \text{black} \,|\, \text{ any } A_N, B_N, C_N) = 1/2$$

And now the other shoe drops as you knew it must. This proper application of probability as inferencing **IS NOT** determining the color of the updated cell. It is instead an accurate reflection of an IP's degree of belief in the truth of the statement that the current cell is colored black. The IP relied on one model with just the information, for better or worse, as reflected in all sixteen cells of the joint probability table.

All I can say at this juncture is that, if contrary to this viewpoint, the IP thinks that the probability of 1/2 means that the ontology of the real world (i.e., which the cellular automaton is mimicking) consulted a random number generator rather than a rule, it believes in quantum mechanics rather than causal determinism.

It seems in this case that the IP is rather favorably disposed to believe that Nature has given up on causal determinism at some level (perhaps due to exhaustion or boredom) and resorts to randomness. What kind of *scientific* mind set is this?

9.3 Relevant Cell Colors Are Unknown

An IP's missing information may show up in other ways. This notion of "missing information" must be rigorously defined, and it is. Any joint probability table with numerical assignments *gratis* some model is inserting only that "information" that remains after leaving out the missing information.

Any missing information is rigorously defined by Shannon's entropy function. Any active information is just as rigorously defined by the average of a constraint function. Implementing this procedure as a variational optimization problem is the essence of the maximum entropy principle that makes numerical assignments to the probabilities as conditioned on the information resident under some model.

Supplemental Exercise 9.3.1: What if the IP doesn't know the colors of all three relevant cells at the previous time step?

Solution to Supplemental Exercise 9.3.1:

Suppose that the IP only knows the color of the left neighboring cell at the previous time step. Can it still make an inference about the color of the current cell under consideration? The examples in Chapter Nine of Volume I showed that departure from strictly deterministic CA was no cause for alarm as far as inferencing was concerned.

The above question about not knowing relevant details necessary for the smooth operation of a CA represents a stripped–down abstract version of the essential characteristic distinguishing an inference from a deduction. Cellular automata are deterministic systems and so are "brittle." If any of the necessary ingredients for a deduction go missing, then everything goes up in flames. The current cell can only be updated if the colors of all three neighboring cells at the previous time step are known. Or, in logic, nothing about a statement A can be deduced by an implication if only B is known.

But inference doesn't suffer from the "brittleness" inherent in deductions. A quantitative measure of the degree of belief in the truth of some statement is still possible under an inference even when the necessary ingredients for the deduction are lacking.

Following along from Chapter Eight, it is easy to set up a 16 cell joint probability table that enforces determinism for, say, an elementary CA operating according to Rule 110. It is always the case that the probability of the color for the current cell under consideration is 1 or 0 when all the necessary ingredients for the deduction are known. Three black cells lead to a white cell with certainty,

$$P(B_{N+1} \,|\, A_N, B_N, C_N, \text{ Rule 110 }) = 0$$

Probability theory will generalize and reproduce the deductions that logic, or the smooth functioning of the ECA, would arrive at on its own.

In order the answer the question posed above, probability theory has a very simple rule, the **Sum Rule**, that *marginalizes* over everything that is not known. Suppose, then, that the cell to the left at the previous time step $\overline{A}_N$ is known to be white, but the colors of the other two cells B_N or $\overline{B}_N$ and C_N or $\overline{C}_N$ are not known. What is the probability that the current cell under consideration, $\overline{B}_{N+1}$, is white?

Applying Bayes's Theorem, we have,

$$\begin{aligned} P(\overline{B}_{N+1} \,|\, \overline{A}_N, \text{ Rule 110}) &= \frac{P(\overline{B}_{N+1}, \overline{A}_N \,|\, \text{ Rule 110})}{P(\overline{A}_N \,|\, \text{ Rule 110})} \\ &= \frac{P(\overline{B}_{N+1}, \overline{A}_N \,|\, \text{ Rule 110})}{P(\overline{B}_{N+1}, \overline{A}_N \,|\, \text{ Rule 110}) + P(B_{N+1}, \overline{A}_N \,|\, \text{ Rule 110})} \end{aligned}$$

Invoking the **Sum Rule**,

$$P(\overline{B}_{N+1}, \overline{A}_N \,|\, \text{ Rule 110}) = \sum_{B_N, C_N}^{4} P(\overline{B}_{N+1}, \overline{A}_N, B_N, C_N \,|\, \text{ Rule 110})$$

$$P(B_{N+1}, \overline{A}_N \,|\, \text{ Rule 110}) = \sum_{B_N, C_N}^{4} P(B_{N+1}, \overline{A}_N, B_N, C_N \,|\, \text{ Rule 110})$$

At this juncture, consult the joint probability table for Rule 110 in Figure 9.1 of Volume I to determine the four cells that are being summmed over,

$$\sum_{B_N,C_N}^{4} P(\overline{B}_{N+1}, \overline{A}_N, B_N, C_N \mid \text{ Rule 110}) = \text{Cell 13} + \text{Cell 14} + \text{Cell 15} + \text{Cell 16}$$

$$= 0 + 0 + 0 + 1/8$$

$$\sum_{B_N,C_N}^{4} P(B_{N+1}, \overline{A}_N, B_N, C_N \mid \text{ Rule 110}) = \text{Cell 5} + \text{Cell 6} + \text{Cell 7} + \text{Cell 8}$$

$$= 1/8 + 1/8 + 1/8 + 0$$

Substituting these results back into Bayes's Theorem, we have the answer of a probability equal to 1/4 that the current cell is colored white given that one of the relevant cells was white and the other two unknown, and in addition that the numerical assignments to the joint probability table ensued from the information in the logic function of Rule 110.

$$P(\overline{B}_{N+1} \mid \overline{A}_N, \text{ Rule 110}) = \frac{1/8}{1/8 + 3/8}$$

$$= 1/4$$

Go back and examine Rule 110 in Figure 3.2 of Volume I. You will observe that in the four cases where the left cell is white, the current cell is colored white only once. This is a pleasant reaffirmation of what intuition would have suggested. Despite this, do not suppose that I have become an advocate for a frequency definition of probability!

9.4 Information about Rules

As first broached in section 9.3.2 of Volume I, the IP may have some uncertainty about exactly which rule is behind the operation of a cellular automaton. As a matter of fact, most people would consider this to be the core issue of the scientific enterprise. We see the world operating in a certain way, but have no clue as to who is behind the curtain pulling all the strings.

Drawing from its fundamental conceptual straightjacket as voluntarily imposed by probability theory, the IP can be said to possess some information about all the CA rules. This information, or equally well from the ying and yang of it all, the amount of *missing* information about the rules, is encapsulated in a probability distribution over all models. The familiar symbolic expression is $P(\mathcal{M}_k)$.

Initial insight into these matters is hard won. I prefer to poke my toe ever so gingerly into potentially deep waters. So to that end, let's restrict the number of

models under consideration to just $\mathcal{M} = 3$. Nevertheless, even this modest venture is an improvement over the two models of section 9.3.2. Retain the two rules, Rule 110 and Rule 30, examined there and augment the space of models with a third rule, Rule 150. Choosing these models means we won't have to go through the trouble of generating new joint probability tables.

In the numerical example studied in Exercise 9.5.16 of Volume I, we applied this particular instantiation for the general posterior predictive formula,

$$P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, \overline{C}_N) = \sum_{k=1}^{\mathcal{M}} P(\overline{B}_{N+1} \,|\, A_N, \overline{B}_N, \overline{C}_N, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, A_N, \overline{B}_N, \overline{C}_N)$$

In other words, the IP wanted to compute the probability for the current cell to be colored white given that the three relevant cells at the previous step were colored black, white, and white. To achieve this goal, the IP averaged over the determined CA output dictated by two models, Rule 110 and Rule 30. This average was taken with respect to the *posterior* probability distribution over model space.

Supplemental Exercise 9.4.1: Recall where we have already constructed a joint probability table for the third model, Rule 150.

Solution to Supplemental Exercise 9.4.1:

Figure 8.2 in these Supplemental Exercises provides a joint probability table that enforces the deterministic output from a Rule 150 cellular automaton. The joint probability tables implementing Rules 110 and 30 appear in Figures 9.1 and 9.2 of Volume I.

Supplemental Exercise 9.4.2: Set up the posterior predictive formula for the new example.

Solution to Supplemental Exercise 9.4.2:

For a change of pace, suppose that the IP wants to compute the probability that the current cell is colored black given that the three relevant cells at the previous step were colored white, black, and black. The model space consists now of $\mathcal{M} = 3$ models.

$$P(B_{N+1} \,|\, \overline{A}_N, B_N, C_N) = \sum_{k=1}^{3} P(B_{N+1} \,|\, \overline{A}_N, B_N, C_N, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, \overline{A}_N, B_N, C_N)$$

Supplemental Exercise 9.4.3: How does the second term on the right hand side get expanded?

Solution to Supplemental Exercise 9.4.3:

The second term is the posterior probability distribution for the three models. Its expansion according to the formal manipulation rules of probability theory looks like,

$$P(\mathcal{M}_k \,|\, \overline{A}_N, B_N, C_N) = \frac{P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{\sum_{k=1}^{3} P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}$$

The conditional probability $P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k)$ is itself expanded via,

$$P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k) = P(B_{N+1}, \overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k) + P(\overline{B}_{N+1}, \overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k)$$

We can now compute the posterior probability for each of the three models. Matching up $\mathcal{M}_1$ with Rule 30, $\mathcal{M}_2$ with Rule 110, and $\mathcal{M}_3$ with Rule 150, we find first that the conditional probability is always 1/8 no matter which model is conditioned on,

$$\begin{aligned} P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k) &= P(B_{N+1}, \overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k) + P(\overline{B}_{N+1}, \overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k) \\ &= 0 + 1/8 = 1/8 \qquad \text{conditioned on Rule 30} \\ &= 1/8 + 0 = 1/8 \qquad \text{conditioned on Rule 110} \\ &= 1/8 + 0 = 1/8 \qquad \text{conditioned on Rule 150} \end{aligned}$$

What is the prior probability $P(\mathcal{M}_k)$ of each model? The IP adopts the stance that it knows nothing permitting it to distinguish among the numerical assignments seen in the three joint probability tables under the information provided by the three models. It asserts that its degree of belief in the truth of the assignments made by Rule 30 is the same as for Rule 110 and Rule 150. It therefore follows Laplace's advice under such a state of knowledge about causes to assign equal probability to all three models. Thus, we have for the "totally uninformed IP" a prior probability of $P(\mathcal{M}_k) = 1/3$.

Everything is now in place for computing the posterior probability distribution over all three models, and thus clearing up what the second term in the posterior predictive formula will look like.

$$
\begin{aligned}
P(\mathcal{M}_k \,|\, \overline{A}_N, B_N, C_N) &= \frac{P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{\sum_{k=1}^{3} P(\overline{A}_N, B_N, C_N \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)} \\
&= \frac{1/8 \times 1/3}{(1/8 \times 1/3) + (1/8 \times 1/3) + (1/8 \times 1/3)} \\
&= 1/3
\end{aligned}
$$

The posterior probability of each of the three models is exactly the same as its prior probability. Knowing the colors of the cells at the previous time step could not change an IP's degree of belief in the truth of a model's assignment to a joint probability table.

Pondering this for a moment (granting the IP the capacity to ponder), the IP reasons that every single rule must be defined by knowing the colors at the previous time step. So, conditioning on such a fact (three fixed cell locations and their colors) which must be true for every rule, and furthermore a fact which cannot discriminate among the rules doesn't help out when the IP wants to update its degree of belief in the rules.

Supplemental Exercise 9.4.4: Compute the probability for the current cell to be black.

Solution to Supplemental Exercise 9.4.4:

Despite the results from the last exercise, the IP proceeds to compute, through the auspices of the posterior predictive formula, that the current cell is colored black with probability 2/3 when the three relevant cells at the previous time step were white, black, and black.

$$
\begin{aligned}
P(B_{N+1} \,|\, \overline{A}_N, B_N, C_N) &= \sum_{k=1}^{3} P(B_{N+1} \,|\, \overline{A}_N, B_N, C_N, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, \overline{A}_N, B_N, C_N) \\
&= (1 \times 1/3) + (1 \times 1/3) + (0 \times 1/3) \\
&= 2/3
\end{aligned}
$$

Two rules, Rule 30 and Rule 110, produce a black cell when the three relevant cells at the previous step are white, black, and black. The final rule from the model space of three rules, Rule 150, produces a white cell.

Supplemental Exercise 9.4.5: What's the point of going to all the trouble of computing these probabilities?

Solution to Supplemental Exercise 9.4.5:

The reason why we do these computations is not for the pleasure of staring at some numbers, but rather in the hope of achieving some real *insight*.

We want to subject these probability formulas, as they have been derived from the formal manipulation rules, to every imaginable "stress" test of our devising. If they should seem to fail any one of these tests, a serious review is in order to see if one can rationalize a justification, or whether, as Jaynes was wont to say, our "intuition" might be in need of some tinkering.

Based on the last exercise, does it now seem plausible that if the IP were to consider a model space made up of a hundred rules, and 99 of these rules output a current black cell given a white, black, and black cell while the one remaining rule output a white cell, then the probability of seeing a black cell is 99/100?

It would seem to be a high priority to subject the posterior predictive formula to as many tough tests as we can devise. After all, this formula is telling an IP how much it can believe in the future occurrence of an event given how often it happened in the past.

$$P(A_{N+1} \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_{N+1} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, A_1, A_2, \cdots, A_N)$$

Supplemental Exercise 9.4.6: Is it possible for an IP's state of knowledge to stay the same even after collecting some data? Use the elementary CA for an illustration.

Solution to Supplemental Exercise 9.4.6:

Within the realm of the 256 elementary CA, what would an IP's state of knowledge be concerning the appearance of a black cell at time step $N+1$ given that it was black at the last time step and its two nearest neighbors were white and black?

The IP doesn't know which rule is governing the evolution of the CA, but it has collected four data points all leading to an updated white cell. The causal factors consisting of the three relevant cell colors at the previous time step were: (1) BBB, (2) BBW, (3) BWB, and (4) BWW, all leading to a white cell.

As we saw in the previous exercise, the template for the probability of the next statement conditioned on all of the data leads to a comparable probability for an updated cell color within the realm of the elementary CA,

$$P(B \,|\, WBB, \mathcal{D}) = \sum_{k=1}^{256} P(B \,|\, WBB, \mathcal{M}_k)\; P(\mathcal{M}_k \,|\, \mathcal{D})$$

Now, before any data, this same formula leads to eminently sensible result that the probability of a black cell is 1/2.

$$P(B\,|\,WBB) = \sum_{k=1}^{256} P(B\,|\,WBB, \mathcal{M}_k)\; P(\mathcal{M}_k) = 1/2$$

Exactly half of the rules lead to a probability of 1 and exactly half of the rules lead to a probability of 0 for $P(B\,|\,WBB, \mathcal{M}_k)$. Since the IP is completely uninformed about which rule is governing the evolution of the CA, it attaches a prior probability of,

$$P(\mathcal{M}_k) = 1/256$$

to every rule. After the summation, we are left with,

$$P(B\,|\,WBB) = 128/256 = 1/2$$

But there is hope that collecting some data on the operation of the CA will allow the IP to update its state of knowledge about the probability of the color for the cell scheduled to be updated. Consider the following argument on how the IP can immediately reduce the model space based on the data.

The first row of Table 9.1 shows the conventional layout of three variables as the arguments for some logic function. The T and F are converted to black and white in the next row. The data told us that the first four configurations of the relevant cells all led to white. The coefficient must then be 0 for the building–block functions 2^7 through 2^4 for any rule.

Table 9.1: *How the data pared down the space of potential rules acting as models.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
BBB	BBW	BWB	BWW	WBB	WBW	WWB	WWW
W	W	W	W	$\star$	$\star$	$\star$	$\star$
0	0	0	0	$\star$	$\star$	$\star$	$\star$
2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
0	0	0	0	$\star$	$\star$	$\star$	$\star$

Therefore, the model space of rules has been pared down to just 16 rules after conditioning on the data. Expanding the resulting base 2 numbers,

$$(0 \times 2^7) + (0 \times 2^6) + (0 \times 2^5) + (0 \times 2^4) + (\star \times 2^3) + (\star \times 2^2) + (\star \times 2^1) + (\star \times 2^0)$$

we see that only Rules 0 through 15 are still in the running as potential models.

Thus,

$$P(B\,|\,WBB,\mathcal{D}) = \sum_{k=1}^{16} P(B\,|\,WBB,\mathcal{M}_k)\,P(\mathcal{M}_k\,|\,\mathcal{D})$$

The remaining 16 models have an equal probability of,

$$P(\mathcal{M}_k\,|\,\mathcal{D}) = 1/16$$

If we literally look at the output from all 16 ECA running according to Rules 0 through 15 when the relevant cells are colored white, black, and black, we observe that the first eight rules produce a white cell and the last eight rules produce a black cell.

We are right back where we started since,

$$\sum_{k=1}^{16} P(B\,|\,WBB,\mathcal{M}_k)\,P(\mathcal{M}_k\,|\,\mathcal{D}) = \overbrace{(0\times 1/16)+\cdots}^{8\text{ terms}} + \overbrace{(1\times 1/16)+\cdots}^{8\text{ terms}}$$

$$P(B\,|\,WBB,\mathcal{D}) = 8/16 = 1/2$$

The IP was unable to update its state of knowledge about the degree of belief in the truth of the statement that the cell at time step $(N+1)$ would be colored black even though some data pared down the model space quite significantly.

The *Mathematica* function that enables us to look at a collection of rules and directly assess the color of an updated cell based on the colors of three relevant cells at the previous time step is **RulePlot[ ]**. I wrote a simple program so that I could scan down the list of the first sixteen rules to see that the first eight rules produced a white cell and the last eight rules produced a black cell for the situation of WWB.

```
Column[Table[{i, RulePlot[CellularAutomaton[i],
                                    Frame → False]}, {0, 15}]]
```

Chapter 10

Logic Puzzles

10.1 The Halloween Party Logic Puzzle

We will try to confirm and expand the results given in Chapter Ten of Volume I for the Halloween Party logic puzzle. These supplemental exercises will be more concise and systematic since we will use *Mathematica* right from the start.

The lengthier, more discursive, and rambling discussions of Chapter Ten will be supported by the *Mathematica* evaluations in the upcoming exercises. I make no apologies for rambling discussions when first broaching some important foundational concept. Conciseness and rigor may enter at some later stage.

More importantly, the investigation begun in Chapters Two and Seven into logic puzzles, tautologies, and the implications for Bayes's Theorem will be elaborated on in the context of the Halloween Party logic puzzle.

The most important conceptual discovery is greater insight into those occasions when it is forbidden to use Bayes's Theorem. As was first expounded earlier in the supplemental exercises for Chapter Two, Bayes's Theorem can not be used in the following instance: A model asserts that some joint statement can not happen, and yet the truth of that joint statement is nonetheless asserted together with the model to the right of the conditioned upon symbol.

It is axiomatic in Classical Logic that everything and nothing can be proven on a contradiction. Therefore, steps must be taken to excise, as Hofstadter says, "this global instantaneous cancer." Probability theory encompasses the same idea within Bayes's Theorem by disallowing contradictory statements to be asserted as true. The subtlety is that one of the contradictory statements may be hidden away within the model. We will discover where this happens in the Halloween Party logic puzzle if we should try to use Bayes's Theorem to make an invalid inference.

Supplemental Exercise 10.1.1: Recall the premises of the Halloween Party puzzle and then translate into *Mathematica* symbolic expressions.

Solution to Supplemental Exercise 10.1.1:

In Exercise 10.4.1 of Volume I, the premises of the Halloween Party puzzle were laid out with their symbolic logic expressions. Repeating these, and now adding the equivalent *Mathematica* expressions,

First premise. "If Alice goes, then Ben won't go and Charlie will."
$A \to [\,\overline{B} \wedge C\,]$
```
Implies[a, And[Not[b], c]]
```

Second premise. "If Ben and Diane go, then either Alice or Charlie (but not both) will go."
$[\,B \wedge D\,] \to [\,(\overline{A} \wedge C) \vee (A \wedge \overline{C}\,]$
```
Implies[And[b, d], Or[And[Not[a], c],
                      And[a, Not[c]]]]
```

Third premise. "If Charlie goes and Ben does not, then Diane will go, but Alice will not."
$[\,\overline{B} \wedge C\,] \to [\,\overline{A} \wedge D\,]$
```
Implies[And[Not[b], c], And[Not[a], d]]
```

Supplemental Exercise 10.1.2: Confirm the DNF for the conjunction of the premises.

Solution to Supplemental Exercise 10.1.2:

Use **`BooleanTable[]`** to find all six terms in the fully expanded DNF. Evaluating,

```
BooleanTable[And[Implies[a, And[Not[b], c]],
     Implies[And[b, d], Or[And[Not[a], c], And[a, Not[c]]]],
     Implies[And[Not[b], c], And[Not[a], d]]]]
```

returns a list of 16 elements consisting of six **`True`** and ten **`False`** elements.

The first eight elements are all **`False`** with the next eight elements looking like this **`{..., True, True, False, True, True, False, True, True}`**. Matching up the six **`True`** elements with **`Tuples[{True, False}, 4]`**, construct the six terms of the fully expanded DNF as,

$$\overline{A}BCD \vee \overline{A}BC\overline{D} \vee \overline{A}B\overline{C}\,\overline{D} \vee \overline{A}\,\overline{B}CD \vee \overline{A}\,\overline{B}\,\overline{C}\,D \vee \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}$$

Supplemental Exercise 10.1.3: Confirm the placement of 0s in the joint probability table under the information of this logic model.

Solution to Supplemental Exercise 10.1.3:

Figure 10.3 of Volume I showed the joint probability table with the probability assignments of 0 and 1/6 placed into all 16 cells. From the results of the last exercise, the six cells corresponding to the six terms in the DNF expansion will contain non–zero terms.

These are cells 9, 10, and 12 in the lower left half of the table, and cells 13, 15, and 16 in the lower right half. Cell 9 is labeled as $\overline{A}BCD$, ..., through cell 16 labeled as $\overline{A}\,\overline{B}\,\overline{C}\,\overline{D}$. The non–zero probabilities must all be in the lower half of the joint probability table because all six terms in the DNF begin with $\overline{A}$.

The remaining ten cells of the joint probability table, cells 1 through 8, together with cells 11 and 14, must contain a probability assignment of 0. These are the ten terms in the orthonormal expansion of the logic expression for the premises with a coefficient that evaluated to F.

Supplemental Exercise 10.1.4: What does *Mathematica* return as the DNF?

Solution to Supplemental Exercise 10.1.4:

Run **BooleanConvert[ ]** with the argument the conjunction of the premises, the same as for **BooleanTable[ ]**, to find out how *Mathematica* reports back the DNF. *Mathematica* will always try to "simplify" what I have called the full expansion DNF.

I don't particularly like this because I want to match up each individual term in the DNF with the individual cells of the joint probability table in order to locate the placement of the zero and non–zero probability assignments. I have to expand "by hand" the "simplified" DNF answer returned by *Mathematica*,

```
BooleanConvert[And[Implies[a, And[Not[b], c]],
     Implies[And[b, d], Or[And[Not[a], c], And[a, Not[c]]]],
     Implies[And[Not[b], c], And[Not[a], d]]]] // FullForm
```

returns only four terms as the DNF,

```
Or[And[Not[a], b, c], And[Not[a], Not[b], Not[c]],
           And[Not[a], c, d], And[Not[a], Not[c], Not[d]]]
```

or $\overline{A}BC \vee \overline{A}\,\overline{B}\,\overline{C} \vee \overline{A}CD \vee \overline{A}\,\overline{C}\,\overline{D}$ the disjunction of four terms in our logic notation.

Each one of these four terms with only three variables will be expanded into two terms with four variables by the **Sum Rule** leading to a total of eight terms. Therefore, two of the terms must be duplicates to get back to the required six terms in the full expansion DNF.

Carrying out each application of the **Sum Rule**, and marking with a $\surd$ the duplicates, we are able to recover our original six terms,

$$\overline{A}BC \rightarrow \overbrace{\overline{A}BCD}^{\surd} \vee \overline{A}BC\overline{D}$$

$$\overline{A}\,\overline{B}\,\overline{C} \rightarrow \overline{A}\,\overline{B}\,\overline{C}D \vee \overbrace{\overline{A}\,\overline{B}\,\overline{C}\,\overline{D}}^{\surd}$$

$$\overline{A}CD \rightarrow \overbrace{\overline{A}BCD}^{\surd} \vee \overline{A}\,\overline{B}CD$$

$$\overline{A}\,\overline{C}\,\overline{D} \rightarrow \overline{A}B\overline{C}\,\overline{D} \vee \overbrace{\overline{A}\,\overline{B}\,\overline{C}\,\overline{D}}^{\surd}$$

Supplemental Exercise 10.1.5: Does a tautology exist between the premises of the Halloween Party puzzle and the conclusion that Alice will not be attending the party?

Solution to Supplemental Exercise 10.1.5:

Back in Chapter Two of these Supplemental Exercises, what we originally thought might be a tautology between the premises and the conclusion in the budget, taxes, and prices puzzle turned out to be wrong. But even more revealing was that at those variable settings where the tautology failed, Bayes's Theorem had a concomitant problem in trying to condition on contradictory statements.

We discovered in examining that logic puzzle that if the information under the model implementing the premises was assumed to be true, and in addition, variable settings that the model said had a probability of 0 were also assumed true, then the result from Bayes's Theorem was undefined because there was a division by 0 in the denominator. This was an example of the aforementioned conditioning on contradictory statements that Bayes's Theorem forbids.

Does a similar situation exist for the Halloween Party puzzle? We know from the joint probability table that $P(\overline{A}\,|\,\mathcal{M}_k) = 1$, that is, Alice will not attend the party. Checking whether a tautology exists between the stated premises and the conclusion that Alice will not attend shows that it fails. Try checking for the tautology with the simplified DNF just found as $\overline{A}BC \vee \overline{A}\,\overline{B}\,\overline{C} \vee \overline{A}CD \vee \overline{A}\,\overline{C}\,\overline{D}$,

```
TautologyQ[Equivalent[Or[
  And[Not[a], b, c], And[Not[a], Not[b], Not[c]],
  And[Not[a], c, d], And[Not[a], Not[c], Not[d]]], Not[a]]]
```

Sure enough, *Mathematica* evaluates this as **`False`**.

Therefore, we are going to have some issues with contradictory statements if we try to use Bayes's Theorem.Where does the tautology fail?

Running a **`BooleanTable[ ]`** on the same expression that was the argument for **`TautologyQ[ ]`** above, we find that it fails at the two variable settings of $\overline{A}B\overline{C}D$ and $\overline{A}\,\overline{B}C\overline{D}$.

Now this result makes complete sense if we are attuned to casting our logic puzzles into inferential problems with an associated joint probability table. These two problematical variable settings for the tautology happen to be cell 11 and cell 14 where a 0 is located.

If we ever have a denominator in Bayes's Theorem that includes a $P(\overline{A}B\overline{C}D)$ or a $P(\overline{A}\,\overline{B}C\overline{D})$ together with any other assigned probability of 0, we are going to end up with Bayes's Theorem as undefined. We have tried to condition on the information resident in a model that says that something cannot happen, and then in the same breath, assume as true that very thing that cannot happen. Bayes's Theorem rightfully balks when we attempt to reason conditioned on contradictory statements.

Supplemental Exercise 10.1.6: Apply Bayes's Theorem to calculate the probability that Alice will attend the Halloween party under some stated conditions of her friends's decisions to go or not go.

Solution to Supplemental Exercise 10.1.6:

What is the probability that Alice will attend given that Ben will attend, but Charlie will not attend, and Diane will attend? Invoking Bayes's Theorem conditioned on the information $\mathcal{M}_k$ provided in the puzzle,

$$P(A\,|\,B\overline{C}D,\mathcal{M}_k) = \frac{P(AB\overline{C}D\,|\,\mathcal{M}_k)}{P(AB\overline{C}D\,|\,\mathcal{M}_k) + P(\overline{A}B\overline{C}D\,|\,\mathcal{M}_k)}$$

$$= \frac{0}{0+0} \qquad \textbf{Undefined!}$$

The 0 probability in cell 11 shows up in the denominator together with the 0 probability in cell 3. Bayes's Theorem is undefined. We are asserting in our symbolic probability expression that model $\mathcal{M}_k$ is true. At the same time, the expression is asserting that $P(B\overline{C}D\,|\,\mathcal{M}_k)$ is also true. But $P(B\overline{C}D\,|\,\mathcal{M}_k)$ has a probability of 0 under model $\mathcal{M}_k$. It cannot happen. The IP is trying to condition on contradictory statements.

We have the other case that asks whether Alice will attend given that Ben and Diane won't attend, but Charlie will.

$$P(A \,|\, \overline{B}C\overline{D}, \mathcal{M}_k) = \frac{P(A\overline{B}C\overline{D} \,|\, \mathcal{M}_k)}{P(A\overline{B}C\overline{D} \,|\, \mathcal{M}_k) + P(\overline{A}\,\overline{B}C\overline{D} \,|\, \mathcal{M}_k)}$$

$$= \frac{0}{0+0} \qquad \textbf{Undefined!}$$

The 0 probability in cell 14 shows up in the denominator together with the 0 probability in cell 6. Bayes's Theorem is undefined. We are asserting in our symbolic probability expression that model $\mathcal{M}_k$ is true. At the same time, the expression is asserting that $P(\overline{B}C\overline{D} \,|\, \mathcal{M}_k)$ is also true. But $P(\overline{B}C\overline{D} \,|\, \mathcal{M}_k)$ has a probability of 0 under model $\mathcal{M}_k$. It cannot happen. The IP is once again trying to condition on contradictory statements.

These were the two problematic cases pointed out to us by the failure of the tautology. Is Bayes's Theorem similarly undefined for this next set of circumstances for everybody's attendance?

What is the probability that Alice will attend given that Ben will not attend, but both Charlie and Diane will?

$$P(A \,|\, \overline{B}CD, \mathcal{M}_k) = \frac{P(A\overline{B}CD \,|\, \mathcal{M}_k)}{P(A\overline{B}CD \,|\, \mathcal{M}_k) + P(\overline{A}\,\overline{B}CD \,|\, \mathcal{M}_k)}$$

$$= \frac{0}{0+1/6}$$

$$= 0$$

Here Bayes's Theorem works just fine. A perfectly legitimate probability of 0 is calculated for the given conditions; Alice will not attend the party. $P(\overline{B}CD \,|\, \mathcal{M}_k)$ was not impossible under the model. Therefore, this particular joint statement could be asserted as true along with the truth of the model. The IP was not attempting to condition on contradictory statements in this case.

Supplemental Exercise 10.1.7: Can we now provide a more complete answer to Exercise 10.4.6 in Volume I?

Solution to Supplemental Exercise 10.1.7:

With the lessons learned from the last few exercises, it is possible to give a better answer to a Chapter Ten exercise that asked the IP to use Bayes's Theorem to verify Alice's non–attendance at the Halloween party under any circumstances.

Inspection of the joint probability table in Figure 10.3 of Volume I tells the IP quite unambiguously that, under the information provided in the premises for the puzzle, the probability of Alice's attendance is 0, $P(A \,|\, \mathcal{M}_k) = 1 - P(\overline{A} \,|\, \mathcal{M}_k) = 0$.

There is nothing wrong with this probability expression. The degree of belief in statement A is conditioned solely on the truth of model $\mathcal{M}_k$. It is certain that Alice will not attend the party.

But what about those two problematic 0s in cells 11 and 14? These 0s appear in the joint probability table under $\overline{A}$ label. Is it therefore FALSE that Alice will NOT attend? This seems to contradict everything we have just said.

This is the point in the conversation where some serious semantic gyrations take place. It is certainly the case that the model places 0s in cells 11 and 14.

However, the joint statements $B\overline{C}D$ and $\overline{B}C\overline{D}$ can never happen under the information from the model! Thus, IT WILL NEVER HAPPEN THAT IT IS FALSE THAT ALICE WILL NOT ATTEND!

We have already proven that in six cases, that is, those cases that don't involve cells 11 and 14 in the denominator of Bayes's Theorem, the IP can condition on any set of circumstances of Alice's friends's attendance or non–attendance. The perfectly legitimate invocation of Bayes's Theorem in these six cases returns the correct probability of Alice's attendance as 0. For example, Exercise 10.4.6 returned the result from Bayes's Theorem that it was certain that Alice would not attend given that Ben did attend, but Charlie and Diane did not.

The two cases that do involve conditioning on the truth of the joint statements indexed by cells 11 and 14 can never happen under the information inserted into the joint probability table that also assumes the truth of model $\mathcal{M}_k$. So, placing either $B\overline{C}D$ or $\overline{B}C\overline{D}$ to the right of the conditioned upon symbol, along side this particular model, is not allowed. The rules of probability theory forbid the use of Bayes's Theorem on the supposition of contradictory statements.

10.2 Clausal Form Logic

Brown presented the Halloween Party puzzle in his Chapter 6 as an example of what he called *clausal form logic*. Moreover, the subject matter of his Chapter 6 was something called *syllogistic reasoning*. Now, this topic is near and dear to our hearts because it uses Classical Logic in essentially the same way we have presented it in Volume I to solve "propositional problems."

I was honestly intrigued by this approach (and I would heartily recommend its consideration by anyone), but after a fair amount of diligent effort on my part, I reluctantly came to the conclusion that, in the end, inferencing through probability theory was much easier to understand, and more generalizable as well.

In my opinion, the results Brown obtains with great difficulty through syllogistic reasoning can be duplicated with far less effort and a far more comprehensive grasp of what is actually going on through a general attack leveraging inferencing and probability theory. That was why I chose his two logic puzzles for my Chapter Ten

of Volume I to illustrate the formal manipulation rules of probability, and more specifically Bayes's Theorem. My probabilistic solution to the Halloween Party puzzle serves as stark contrast to (what to me) are the mysteries of "syllogistic reasoning."

Nonetheless, Brown's comments on *clausal form logic*, together with his excellent explanation, are worthy of discussion here. Here is a quote from Brown that caught my attention.

> Mechanical theorem–proving and other forms of computer based reasoning require that logical statements be supplied in a standardized language. A format called *clausal form logic* (CFL) ... has been adopted, in one form or another, by most automated reasoning systems. Richards writes of CFL
>
> > It is a very simple type of logic; all its statements are of the form *if something then something*. It lacks the usual apparatus of quantifiers ... and connectives ... which are normally associated with logical systems. Yet, despite this apparent poverty, it has as much power as a logic with all these devices, both to represent statements and to carry out reasoning with them ... This simple if–then format of statements in CFL makes it ideally suited for computer based reasoning systems.

Clausal form logic and its variants were indeed the flavor of the day back in the late 1970s and early 1980s as a rigorous way of implementing artificial intelligence. Rule production systems and expert systems that relied on the exclusive use of *if–then rules* became quite popular.

This particular approach to AI was eclipsed in the mid 1980s by the rediscovery and generalization of artificial neural networks. Such networks disdained any need to rely on logic by breaking up and distributing knowledge over a huge number of vastly interconnected cells, and relying on massive amounts of exposure to training events in order to fine tune the connections in the network. Even the adherents of this approach admitted that shifting the representation of knowledge from, say, logic expressions to widely distributed "atoms" within a network meant that nobody understood any longer how the AI system actually reached its conclusions.

Today, it appears that the overwhelming choice for integrating AI into everything from automated vehicles to expert gaming ability to image identification to chatbots appears to be the latter descendants of these original artificial neural networks (ANNs). The exponential increase in software and hardware power over the past thirty years has allowed for a concomitant increase in the number of connected intervening layers between input and output in these advanced ANNs.

It became feasible to expose these networks to what formerly would have been implausible amounts of data. Every connection could be adjusted in an evolutionary fashion so that tiny incremental improvements not noticeable after 100, 500, or even a 1000 data points led to correct conclusions after 10,000,000 data points.

Even more intriguing, these AI systems can generate their own simulated trials, or more evocatively, play millions upon millions of games to see which strategies pan out and which don't. Today, this combined approach has been given the sobriquet of *deep learning*.

Even though its popularity has waned, I still think implementing AI systems with logic implemented somewhere in the overall system is worth continued investigation. Of course, by logic I mean its generalization through inferencing and probability theory. As is oftentimes the case, *hybrid* systems that take advantage of both approaches should be a topic for research. Maybe the low–level processing from input to some intermediate level needs to be handled through deep learning, at which point a summary of that low level processing could be taken over by logic functions, or probabilistic models.

Supplemental Exercise 10.2.1: Following Brown's prescription, generate a couple of easy examples illustrating the *if–then* format of CFL.

Solution to Supplemental Exercise 10.2.1:

Brown defines a clause as a conditional in the general form,

$$a_1 \cdots a_m \rightarrow b_1 + \cdots + b_n$$

There are m propositions A, B, $\cdots$ $(a_1 \cdots a_m)$ joined by AND on the left hand side of the IMPLIES and n propositions C, D, $\cdots$ $(b_1 + \cdots + b_n)$ joined by OR on the right hand side.

Therefore, the clause $a_1 a_2 \rightarrow b_1$, "If A and B, then C," would be expressed first symbolically as $(A \wedge B) \rightarrow C$, and secondly in *Mathematica* as,

```
Implies[And[a, b], c]
```

The slightly more complicated clause $a_1 a_2 \rightarrow b_1 + b_2$, "If A and B, then C or D" would be expressed first symbolically as $(A \wedge B) \rightarrow (C \vee D)$, and secondly in *Mathematica* as,

```
Implies[And[a, b], Or[c, d]]
```

And to finish up this introduction to Brown's definition, generate an acceptable clause with $m = 2$ and $n = 3$, $a_1 a_2 \rightarrow b_1 + b_2 + b_3$, "If A and B, then C or D or E" would be expressed first symbolically as $(A \wedge B) \rightarrow (C \vee D \vee E)$, and secondly in *Mathematica* as,

```
Implies[And[a, b], Or[c, d, e]]
```

Supplemental Exercise 10.2.2: Apply some typical Boolean operations to these CFL expressions.

Solution to Supplemental Exercise 10.2.2:

Insight into these clausal forms can be gained by first subjecting such expressions to our usual DNF decomposition, and then adding a little twist at the end. Examine the DNF of the first example from the last exercise with,

```
BooleanConvert[Implies[And[a, b], c]]
```

which returns **`Or[Not[a], Not[b], c]`**

The truth table found with,

```
BooleanTable[Implies[And[a, b], c]]
```

returns only one **`False`** at the variable settings $A = T$, $B = T$, and $C = F$ with the remaining seven settings all **`True`**. Thus, the fully expanded DNF would consist of those seven terms with a coefficient of T, $ABC \vee \cdots \vee \overline{A}\,\overline{B}\,\overline{C}$.

But in this case it would be easier to focus the DNF expansion on the one term with the coefficient of F which is $AB\overline{C}$. By the "duality" operation (Exercise 5.9.16 in Volume I), this one term $AB\overline{C}$ is the same as the DNF expression that *Mathematica* returned after complementing each variable, switching the operators AND and OR, and finally switching the coefficient from F to T. Thus, we now see the *Mathematica* result of $\overline{A} \vee \overline{B} \vee C$.

It is always easier, to my way of thinking, to check such answers by reverting back to the joint probability table. Thus, $P(AB\overline{C} \,|\, \mathcal{M}_k) = 0$ because of the coefficient of F associated with this term in the DNF expansion, becomes $P(\overline{A} \vee \overline{B} \vee C \,|\, \mathcal{M}_k) = 1$.

Cell 2 contains a 0 because it indexes $P(AB\overline{C} \,|\, \mathcal{M}_k)$. Cells 5, 6, 7, and 8 index $P(\overline{A} \,|\, \mathcal{M}_k)$, cells 3, 4, 7, and 8 index $P(\overline{B} \,|\, \mathcal{M}_k)$, and cells 1, 3, 5, and 7 index $P(C \,|\, \mathcal{M}_k)$. Removing the duplicated cells 3, 5, 7, and 8, we add up the probabilities in cells 1, 2, 3, 4, 5, 6, 7, and 8 excluding only cell 2 which contains a 0. Thus, we can confirm that,

$$P(AB\overline{C} \,|\, \mathcal{M}_k) = 1 - P(\overline{A} \vee \overline{B} \vee C \,|\, \mathcal{M}_k) = 0$$

Supplemental Exercise 10.2.3: Is there something slightly odd about the DNF expression returned by *Mathematica*?

Solution to Supplemental Exercise 10.2.3:

Working through the last exercise provided another example of why I wish that *Mathematica* would return the fully expanded DNF with all seven terms instead of

the simplified version of **`Or[Not[a], Not[b], c]`**. It doesn't make any difference whether you ask for the disjunctive normal form (the default) as in the above exercise, or the conjunctive normal form by specifying,

```
BooleanConvert[Implies[And[a, b], c], "CNF"]
```

The answer returned is exactly the same! There is no way of filling in any of the available *forms* as an argument to **`BooleanConvert[ ]`** that will return all seven terms with a coefficient of T. This (lack of a) feature continues to be a source of minor frustration for me.

Supplemental Exercise 10.2.4: Apply these lessons to better understand Brown's CFL approach to the Halloween Party puzzle. Do you prefer his approach or my probabilistic flavored approach of Chapter Ten?

Solution to Supplemental Exercise 10.2.4:

Brown fully intended the Halloween Party puzzle to be an example of syllogistic reasoning using CFL. The first premise in the puzzle was,

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

for which he gave the equation $AB + AC' = 0$. We reproduce this result by first finding the DNF,

```
BooleanConvert[Implies[a, And[Not[b], c]] // FullForm
```

which returns **`Or[Not[a], And[Not[b], c]]`** or $\overline{A} \vee (\overline{B} \wedge C)$ recognizable as the alternative form for an implication.

Now, we implement the duality operations by switching the operators $\vee$ and $\wedge$, complementing all variables, and realizing that instead of T from the DNF, this new dual expression has the truth value of F.

$$[\,\overline{A} \vee (\overline{B} \wedge C) = T\,] \xrightarrow{\text{dual}} [\,A \wedge (B \vee \overline{C}) = F\,]$$

The expression on the right hand side will be operated on by the formal rules of Boolean manipulation to yield Brown's expression,

$$[\,A \wedge (B \vee \overline{C}) = F\,] \equiv [\,AB \vee A\overline{C} = F\,] \equiv [\,AB + AC' = 0\,]$$

Do exactly the same thing to the second premise, (with apologies to Douglas Hofstadter),

$$BD \to \overline{A}C \vee A\overline{C}$$

for which Brown gave the equation $BD(A'C' + AC) = 0$. We are able to reproduce this result by again finding the DNF,

```
BooleanConvert[Implies[And[b, d], Or[And[Not[a], c],
                        And[a, Not[c]]]]] // FullForm
```

which returns,

```
Or[And[a, Not[c], And[Not[a], c]], Not[b], Not[d]
```

or $A\overline{C} \vee \overline{A}C \vee \overline{B} \vee \overline{D}$.

The duality operations produce a somewhat more complicated expression this time,

$$\begin{aligned}
[\, A\overline{C} \vee \overline{A}C \vee \overline{B} \vee \overline{D} = T \,] &\xrightarrow{\text{dual}} [\, (\overline{A} \vee C) \wedge (A \vee \overline{C}) \wedge B \wedge D = F \,] \\
&= BD \wedge (\overline{A}A \vee \overline{A}\,\overline{C} \vee CA \vee C\overline{C}) \\
&= BD \wedge (\overline{A}\,\overline{C} \vee AC) \\
&= [\, BD\,(A'C' + AC) = 0 \,]
\end{aligned}$$

Everything proceeds just as before for the third and final premise,

$$\overline{B}C \to \overline{A}D$$

for which Brown gave the equation $AB'C + B'CD' = 0$. Reproduce this result by asking *Mathematica* for the DNF

```
BooleanConvert[Implies[And[Not[b], c],
                        And[Not[a], d]]] // FullForm
```

which returns,

```
Or[And[Not[a], d], b, Not[c]]
```

or, in the symbolic logic expression notation,

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

Invoking the duality operations enables us to reproduce Brown's equation,

$$\begin{aligned}
[\, (\overline{A} \wedge D) \vee B \vee \overline{C} = T \,] &\xrightarrow{\text{dual}} (A \vee \overline{D}) \wedge \overline{B} \wedge C = F \,] \\
(A \vee \overline{D}) \wedge \overline{B} \wedge C &= (A\overline{B} \vee \overline{B}\,\overline{D}) \wedge C \\
(A\overline{B} \vee \overline{B}\,\overline{D}) \wedge C &= A\overline{B}C \vee \overline{B}C\overline{D} \\
&= [\, AB'C + B'CD' = 0 \,]
\end{aligned}$$

Supplemental Exercise 10.2.5: Is Brown now ready to apply syllogistic reasoning after everything accomplished in the last exercise?

Solution to Supplemental Exercise 10.2.5:

No, he is not. He now has to add all three terms to obtain an equation in 0, and then extracts common factors. He tells us that at this juncture all that we have to do is,

> The Blake canonical form ... is found by multiplying out to obtain an SOP formula, applying consensus repeatedly, and deleting absorbed terms.

I submit to you that it is far easier to travel down the probability path, as I did in Chapter Ten. Construct the joint probability table and insert 0s in all those cell locations wherever the DNF instructs you that a coefficient of F is associated with an orthonormal functional building block. In other words, if an orthonormal building block in Boole's Expansion Theorem, and suppose it to be $A\overline{B}C\overline{D}$, has a coefficient of F, then a 0 is placed into cell 6.

Then, the IP can make any inference it cares to. Brown gives as a "prime clause" ensuing from the complicated syllogistic reasoning procedure just described that $BD \to C$. If Ben and Diane go to the party, then that implies that Charlie will go as well.

But we can much more easily verify this "prime clause" with Bayes's Theorem,

$$\begin{aligned}
P(C \,|\, B, D, \mathcal{M}_k) &= \frac{P(BCD \,|\, \mathcal{M}_k)}{P(BD \,|\, \mathcal{M}_k)} \\
&= \frac{P(ABCD \,|\, \mathcal{M}_k) + P(\overline{A}BCD \,|\, \mathcal{M}_k)}{P(ABCD \,|\, \mathcal{M}_k) + P(\overline{A}BCD \,|\, \mathcal{M}_k) + P(AB\overline{C}D \,|\, \mathcal{M}_k) + P(\overline{A}B\overline{C}D \,|\, \mathcal{M}_k)} \\
&= \frac{0 + 1/6}{0 + 1/6 + 0 + 0} \\
&= 1
\end{aligned}$$

It is certain that Charlie will go to the party if Ben and Diane are going.

Brown presents another prime clause $A \to 0$, Alice will not go to the party. But we have examined this prime clause rather thoroughly from the perspective of probability theory to confirm that $P(A \,|\, \mathcal{M}_k) = 0$.

Chapter 11

Formal Rules for Prediction

11.1 Coherent Arguments

Sometimes there are different ways of approaching a probability calculation. The approach may be supported by different intuitive or even formal arguments one chooses to marshal on behalf of the calculation. But for me *coherency* would seem to demand that all legitimate arguments must, in the end, lead to the same correct final probability.

I would like to demonstrate what I have in mind by this description of coherent arguments with a few examples involving coin tosses. The first approach takes a rather direct route from our basic understanding of Bayes's Theorem. This approach is not quite the one presented in Chapter Eleven of Volume I in order to begin checking on whether coherency has been maintained.

Supplemental Exercise 11.1.1: Calculate the probability for HEADS on the fourth toss of some unknown coin where the coin has already been tossed three times and the results duly recorded.

Solution to Supplemental Exercise 11.1.1:

Suppose that the first three tosses resulted in TAILS, TAILS, and HEADS. The first symbolic probability expression we write down is the conditional probability of obtaining HEADS on the fourth toss,

$$P(A_4 = \text{HEADS} \,|\, A_1 = \text{TAILS}, A_2 = \text{TAILS}, A_3 = \text{HEADS})$$

where the subscript on A indicates the temporal ordering of each coin toss. For this problem, $N = 3$ with $N_1 = 1$, $N_2 = 2$ and $M = 1$ with $M_1 = 1$ and $M_2 = 0$, and $N + M = 4$. Note that we are conditioning on the exact sequence of the data; not just on the data summary of two TAILS and one HEADS in three tosses.

Invoke Bayes's Theorem right at the outset,

$$P(A_4 \,|\, \overline{A}_1, \overline{A}_2, A_3) = \frac{P(A_4, A_3, \overline{A}_2, \overline{A}_1)}{P(A_3, \overline{A}_2, \overline{A}_1)} = \frac{P(A_4, A_3, \overline{A}_2, \overline{A}_1)}{P(A_4, A_3, \overline{A}_2, \overline{A}_1) + P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1)}$$

where the denominator has been expanded by the **Sum Rule**. The primitive mantra for Bayes's Theorem is: *The conditional probability on the left hand side is equal to a joint probability divided by a marginal probability on the right hand side.*

However, each term in the numerator and denominator of Bayes's Theorem is itself another marginal probability over all models $\mathcal{M}_k$.

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1) = \sum_{k=1}^{\mathcal{M}} P(A_4, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k)$$

$$P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1) = \sum_{k=1}^{\mathcal{M}} P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k)$$

Let's immediately transition to the continuous model space for q over the real line from 0 to 1 so that we can substitute an integration over model space for the discrete sum indicated above.

$$\begin{aligned}
P(A_4, A_3, \overline{A}_2, \overline{A}_1) &= \int_0^1 q \times q \times (1-q) \times (1-q) \times C_{\text{Beta}} \times q^{\alpha-1} \times (1-q)^{\beta-1} \, \mathrm{d}q \\
&= C_{\text{Beta}} \times \int_0^1 q^{2+\alpha-1} \times (1-q)^{2+\beta-1} \, \mathrm{d}q \\
&= \frac{1}{30} \\
P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1) &= \int_0^1 (1-q) \times q \times (1-q) \times (1-q) \times C_{\text{Beta}} \times q^{\alpha-1} \times (1-q)^{\beta-1} \, \mathrm{d}q \\
&= C_{\text{Beta}} \times \int_0^1 q^{1+\alpha-1} \times (1-q)^{3+\beta-1} \, \mathrm{d}q \\
&= \frac{1}{20}
\end{aligned}$$

Substituting these numerical values back into Bayes's Theorem, we find that,

$$P(A_4 \,|\, \overline{A}_1, \overline{A}_2, A_3) = \frac{1/30}{1/30 + 1/20} = 2/5$$

The probability for observing HEADS on the fourth toss of an unknown coin after having seen it tossed three times with the results of TAILS on the first toss, TAILS on the second toss, and HEADS on the third toss is 0.40. This result doesn't jar your intuition because you knew that the probability for HEADS on the first toss was 1/2, and after having seen two TAILS and one HEADS in three tosses, lowering the probability for HEADS to 0.40 does not seem unreasonable.

Supplemental Exercise 11.1.2: Pinpoint and fill in the steps I skipped over in the first exercise.

Solution to Supplemental Exercise 11.1.2:

First of all, what gave me the right to go from the expression on the left hand side to the equivalent expression on the right hand side?

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1) \text{ to } \int_0^1 q \times q \times (1-q) \times (1-q) \times C_{\text{Beta}} \times q^{\alpha-1} \times (1-q)^{\beta-1} \, \mathrm{d}q$$

We can expand the first joint probability expression that contains the observables and the models by the **Product Rule**, another one of our constantly used formal manipulation rules,

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k) = P(A_4 \,|\, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k) \times P(A_3 \,|\, \overline{A}_2, \overline{A}_1, \mathcal{M}_k) \; \times$$
$$P(\overline{A}_2 \,|\, \overline{A}_1, \mathcal{M}_k) \times P(\overline{A}_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

The next transformation is important because it is not part of the axiomatic foundation of probability. The only formal manipulation rule that would allow us to transition from, say, $P(A_4 \,|\, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k)$ to $P(A_4 \,|\, \mathcal{M}_k)$ would be to invoke as an assumption that the known outcomes of the previous three coin tosses are irrelevant to a probability assignment at the fourth toss.

The numerical assignment at the fourth toss is seen to depend *only* on the information contained in model $\mathcal{M}_k$. Furthermore, not only at the fourth toss, but at any toss the assigned probability depends only on the model, and not on any previous data. Things depend as well on the initial definition of the state space where the dimension of the state space was defined as $n = 2$, and not an n of some larger number. In this manner, the complicated expression on the right hand side can be simplified to,

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k) = P(A_4 \,|\, \mathcal{M}_k) \times P(A_3 \,|\mathcal{M}_k) \times P(\overline{A}_2 \,|\, \mathcal{M}_k) \times P(\overline{A}_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

Each one of these terms $P(A_t \,|\, \mathcal{M}_k)$ is defined as q for (A_t = HEADS) and $(1-q)$ for (A_t = TAILS). But q and $(1-q)$ must be able to traverse over every conceivable legitimate value of q between 0 and 1. Each model $\mathcal{M}_k$ is itself a *statement* that might be expressed verbally in whatever way is most pleasing. Such a statement might be something like: "Model $\mathcal{M}_k$ asserts that the assigned probability to an occurrence of HEADS is q and the assigned probability to TAILS is $(1-q)$." As an almost trivial consequence, if the IP conditions on the truth of model $\mathcal{M}_k$, we then have,

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k) = q \times q \times (1-q) \times (1-q) \times P(\mathcal{M}_k)$$

Examining this last expression, it is quite interesting that the formal rules have retained a final term $P(\mathcal{M}_k)$ reserved exclusively for a prior probability over these statements claiming to know the true assignments for HEADS and TAILS.

The axiomatic foundation of probability theory does permit the IP to set up a distribution of probabilities over statements indicating the IP's degree of belief that any particular statement is, in fact, true. When some model $\mathcal{M}_k$ asserts that the absolutely true and correct probability for HEADS is 0.75 and the absolutely true and correct probability for TAILS is 0.25, the IP has the right to express its degree of belief that this assertion is true. It might quite legitimately assign a probability of, say, 1/5 as a prior probability for this model's claim, and keep in reserve the remaining probability for its degree of belief in the assertions made by other competing models.

Models are quite clearly very important components in this whole probabilistic enterprise. Their role as convenient mathematical fictions allow information to be inserted into probability distributions! However, as convenient fictions, they must, in the end, be marginalized out of the probability expressions. That is done here in finding the probability solely for the observables A_4 through $\overline{A}_1$.

$$\begin{aligned} P(A_4, A_3, \overline{A}_2, \overline{A}_1) &= \sum_{k=1}^{\mathcal{M}} P(A_4, A_3, \overline{A}_2, \overline{A}_1, \mathcal{M}_k) \\ &= \sum_{k=1}^{\mathcal{M}} q \times q \times (1-q) \times (1-q) \times P(\mathcal{M}_k) \end{aligned}$$

As the number of models grows to cover every small interval over the real line between 0 and 1, the discrete sum converges to an integration over the region from 0 to 1. The discrete prior probability over the $\mathcal{M}$ models converges to a probability density function over q,

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1) = \int_0^1 q \times q \times (1-q) \times (1-q) \times \text{pdf}\,(q)\,\mathrm{d}q$$

The *beta distribution* is the perfect choice for this probability density function over q. When substituted for pdf (q), we have for the numerator and first term in the denominator of Bayes's Theorem,

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1) = \int_0^1 q \times q \times (1-q) \times (1-q) \times C_{\text{Beta}} \times q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q$$

The second term in the denominator of Bayes's Theorem is then,

$$P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1) = \int_0^1 (1-q) \times q \times (1-q) \times (1-q) \times C_{\text{Beta}} \times q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q$$

The constant term C_{Beta} will cancel out in the numerator and denominator, and we are left with,

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1) \propto \int_0^1 q^2 \times (1-q)^2 \times q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q$$

$$\propto \int_0^1 q^{2+\alpha-1}\,(1-q)^{2+\beta-1}\,\mathrm{d}q$$

$$P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1) \propto \int_0^1 (1-q) \times q \times (1-q) \times (1-q) \times q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q$$

$$\propto \int_0^1 q^{1+\alpha-1}\,(1-q)^{3+\beta-1}\,\mathrm{d}q$$

The prior probability over the models is Laplace's prescription for an IP who is uninformed about everything involved in the coin toss. This is a flat or uniform probability density function over all q, or the *beta distribution* with its parameters set to $\alpha = \beta = 1$.

$$P(A_4, A_3, \overline{A}_2, \overline{A}_1) \propto \int_0^1 q^{2+\alpha-1}\,(1-q)^{2+\beta-1}\,\mathrm{d}q$$

$$\propto \int_0^1 q^2\,(1-q)^2\,\mathrm{d}q$$

$$= \frac{2!\,2!}{5!}$$

$$P(\overline{A}_4, A_3, \overline{A}_2, \overline{A}_1) \propto \int_0^1 q^{1+\alpha-1}\,(1-q)^{3+\beta-1}\,\mathrm{d}q$$

$$\propto \int_0^1 q\,(1-q)^3\,\mathrm{d}q$$

$$= \frac{1!\,3!}{5!}$$

Substituting these numerical values back into Bayes's Theorem, we have recovered our required probability for observing HEADS on the fourth coin toss after having observed TAILS, TAILS, and HEADS on the first three tosses,

$$P(A_4 \,|\, \overline{A}_1, \overline{A}_2, A_3) = \frac{1/30}{1/30 + 1/20} = 2/5$$

Supplemental Exercise 11.1.3: What is the distinction to be made in the above exercises whenever the expression for the probability of the data is mentioned?

Solution to Supplemental Exercise 11.1.3:

The previous exercises dealt with the *specific sequence* of the data as in TAILS first toss, TAILS second toss, HEADS third toss, written out succinctly as TTH. Usually we talk about the data $\mathcal{D}$ in summary terms such as N_1 HEADS and N_2 TAILS occurred in N past coin tosses. The same language applies when we talk about the probability for M_1 HEADS and M_2 TAILS in M future tosses, and not about any specific sequence of HEADS and TAILS.

We already know that the probability for any data involving binary outcomes is $P(\mathcal{D}) = \frac{1}{N+1}$. Reverting back to our numerical example, in $N = 3$ coin tosses, the probability of no HEADS, one HEADS, two HEADS, or three HEADS are all the same at $P(\mathcal{D}) = \frac{1}{4}$. And so, obviously, the probability of the data in our example where we observed just one HEADS is 1/4.

But the computed value in the denominator of Bayes's Theorem for our example was, in fact, not 1/4 but rather $1/30+1/20 = 1/12$. Hence, the necessity of drawing the distinction between a specific sequence of the outcomes, and a summary of the outcomes when we use the expression $\mathcal{D}$. There are three ways for the *data* of two TAILS and one HEADS to occur in three tosses of the coin,

$$W(N) = \frac{3!}{1!\,2!} = 3$$

In the numerator of Bayes's Theorem, then, we must take account of $W(N)$ by multiplying our previous probabilities for a specific sequence by $W(N)$. The denominator as just discussed is $P(\mathcal{D}) = 1/4$, so we have,

$$\begin{aligned}
P(M_1 = 1, M_2 = 0 \,|\, N_1 = 1, N_2 = 2) &= \frac{P(M_1, M_2, N_1, N_2)}{P(\mathcal{D})} \\
&= \frac{\int_0^1 q^1\,(1-q)^0\,W(N)\,q^1\,(1-q)^2\ \mathrm{pdf}\,(q)\ \mathrm{d}q}{1/4} \\
&= 4 \times W(N) \times \int_0^1 q^2\,(1-q)^2\ \mathrm{pdf}\,(q)\ \mathrm{d}q \\
&= 4 \times 3 \times \frac{2!\,2!}{5!} \\
&= \frac{2}{5}
\end{aligned}$$

The very satisfying conclusion is that we recapture the same probability whether we are thinking in terms of the actual sequence of outcomes, or just the summary specification expressed as $\mathcal{D}$. But as we shall discover in an upcoming exercise, this is true only when $M = 1$.

11.2 Coherency with Myself?

On page 291, in section 12.5 of Volume I, the explicit formula for the probability of M_1 and M_2 future frequency counts when conditioned on some known data was written out as,

$$P(M_1, M_2 \,|\, N_1, N_2) = C \times \frac{(M_1 + N_1)!\,(M_2 + N_2)!}{M_1!\,M_2!}$$

Were the arguments used in deriving this formula coherent with the arguments in the previous section?

Supplemental Exercise 11.2.1: Using this above formula, do we arrive at the same probability of 2/5 for HEADS on the fourth coin toss given the same data as before?

Solution to Supplemental Exercise 11.2.1:

Reviewing once again, for our current problem, $M = 1$ with $M_1 = 1$ and $M_2 = 0$, and $N = 3$ with $N_1 = 1$ and $N_2 = 2$. The constant term is then,

$$\begin{aligned} C &= \frac{M!\,(N + n - 1)!}{N_1!\,N_2!\,(M + N + n - 1)!} \\ &= \frac{1!\,(3 + 2 - 1)!}{1!\,2!\,(1 + 3 + 2 - 1)!} \\ &= \frac{4!}{2!\,5!} \end{aligned}$$

The second term is,

$$\begin{aligned} \frac{(M_1 + N_1)!\,(M_2 + N_2)!}{M_1!\,M_2!} &= \frac{(1 + 1)!\,(0 + 2)!}{1!\,0!} \\ &= 2!\,2! \end{aligned}$$

Putting these two terms back together, we have,

$$P(M_1, M_2 \,|\, N_1, N_2) = \frac{4!}{2!\,5!} \times 2!\,2! = \frac{2}{5}$$

Apparently, arguments coherent with the other approaches just discussed were used in deriving this formula. Thank goodness these lead once again to the same correct probability!

The original goal of this formula in section 12.5, Volume I, was to show that the general formula when applied to the *next* occurrence was able to reproduce Laplace's **Rule of Succession** formula. It does so in this case,

$$\frac{N_i + 1}{N + 2} = \frac{1 + 1}{3 + 2} = \frac{2}{5}$$

11.3 More Coherent Arguments

Sometimes, I set up an initial joint probability expression looking like this,

$$P(M_1, M_2, N_1, N_2, \mathcal{M}_k)$$

But the **Commutativity** axiom would have just as easily permitted writing down an equivalent expression,

$$P(M_1, M_2, \mathcal{M}_k, N_1, N_2)$$

What are the consequences if this alternative joint probability were taken as the starting point for the derivation?

Supplemental Exercise 11.3.1: Does this different approach also include a coherent argument for arriving at the correct probability?

Solution to Supplemental Exercise 11.3.1:

We would begin with the usual invocation of the **Product Rule**,

$$P(M_1, M_2, \mathcal{M}_k, N_1, N_2) = P(M_1, M_2 \,|\, \mathcal{M}_k, N_1, N_2) \times P(\mathcal{M}_k \,|\, N_1, N_2) \times P(N_1, N_2)$$

followed by the assumed irrelevancy of any past data which merely recorded that HEADS or TAILS did occur as a causal explanation for *why* those HEADS or TAILS occurred,

$$P(M_1, M_2, \mathcal{M}_k, N_1, N_2) = P(M_1, M_2 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k \,|\, N_1, N_2) \times P(N_1, N_2)$$

The causes, for better or worse, are encapsulated within the information provided by some specific model. So goes the justification for the first term on the right hand side above, $P(M_1, M_2 \,|\, \mathcal{M}_k)$.

Notationally, we make use of the following,

$$q \equiv P(A_t = \text{HEADS} \,|\, \mathcal{M}_k)$$

$$(1 - q) \equiv P(A_t = \text{TAILS} \,|\, \mathcal{M}_k)$$

The posterior probability for a model is $P(\mathcal{M}_k \,|\, N_1, N_2)$, while the probability for the data is $P(N_1, N_2) \equiv P(\mathcal{D})$.

If you will pardon me on this occasion, I will jauntily step through what I suppose would be a non–controversial series of transformations,

$$
\begin{aligned}
P(M_1, M_2 \,|\, N_1, N_2) &= \frac{P(M_1, M_2, N_1, N_2)}{P(N_1, N_2)} \\
P(M_1, M_2, N_1, N_2) &= \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, \mathcal{M}_k, N_1, N_2) \\
P(M_1, M_2 \,|\, N_1, N_2) &= \frac{\sum_{k=1}^{\mathcal{M}} P(M_1, M_2, \mathcal{M}_k, N_1, N_2)}{P(N_1, N_2)} \\
P(M_1, M_2, \mathcal{M}_k, N_1, N_2) &= q \times P(\mathcal{M}_k \,|\, N_1, N_2) \times P(N_1, N_2) \\
P(M_1, M_2 \,|\, N_1, N_2) &= \frac{q \times P(\mathcal{M}_k \,|\, N_1, N_2) \times P(N_1, N_2)}{P(N_1, N_2)} \\
P(M_1, M_2 \,|\, N_1, N_2) &= \sum_{k=1}^{\mathcal{M}} q \times P(\mathcal{M}_k \,|\, N_1, N_2) \\
&= \sum_{k=1}^{\mathcal{M}} q \times \frac{P(\mathcal{D} \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)}{P(\mathcal{D})} \\
&= \frac{1}{P(\mathcal{D})} \sum_{k=1}^{\mathcal{M}} q \times P(\mathcal{D} \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \\
&= \frac{1}{P(\mathcal{D})} \sum_{k=1}^{\mathcal{M}} q \times W(N) \times q \times (1-q)^2 \times P(\mathcal{M}_k) \\
&= \frac{1}{P(\mathcal{D})} \times W(N) \times \int_0^1 q \times q \times (1-q)^2 \times \text{pdf}\,(q)\,\mathrm{d}q \\
P(\mathcal{D}) &= \frac{1}{N+1} \\
&= \frac{(N+1)!}{N_1!\,N_2!} \times \int_0^1 q^2 \times (1-q)^2 \times C_{\text{Beta}} \times q^{\alpha-1}(1-q)^{\beta-1}\,\mathrm{d}q \\
&= \frac{(N+1)!}{N_1!\,N_2!} \times C_{\text{Beta}} \times \int_0^1 q^{2+\alpha-1} \times (1-q)^{2+\beta-1}\,\mathrm{d}q \\
&= \frac{(N+1)!}{N_1!\,N_2!} \times 1 \times \frac{\Gamma(3)\,\Gamma(3)}{\Gamma(6)} \\
&= \frac{4!}{1!\,2!} \times \frac{2!\,2!}{5!} \\
&= \frac{2}{5}
\end{aligned}
$$

As always, the final series of steps is predicated on setting the parameters of the *beta distribution* to $\alpha = \beta = 1$ to enforce a uniform distribution over q. All of the causes for why HEADS or TAILS might appear are encapsulated by the information within the models. Each one of these models is given an equal weight through the uniform prior probability.

11.4 Coherency with Jaynes?

Imagine that you had just finished reading Chapter 18 of Jaynes's book [11]. Based on Jaynes's more authoritative stature, you judge it more prudent to employ his formula, rather than mine, for the probability of future events conditioned on some data. Certainly, the formulas being so dissimilar in appearance, it seems plausible that they might give different answers. So you test out Jaynes's combinatorial formula, his Equation (18.24), on my example,

$$P(M_m \mid N_n) = \frac{\binom{n+m}{n}\binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}}$$

Supplemental Exercise 11.4.1: Are my arguments coherent with Jaynes's arguments?

Solution to Supplemental Exercise 11.4.1:

Jaynes's notation of M_m is equivalent to a future occurrence of HEADS where $m = 1$ (one HEADS appears in a future toss) and N_n represents the data where $n = 1$ (one HEADS appeared in the data). Jaynes's N and M are the same as mine so that $N = 3$ and $M = 1$.

$$\begin{aligned} P(\text{HEADS} \mid \mathcal{D}) &= \frac{\binom{1+1}{1}\binom{3+1-1-1}{3-1}}{\binom{3+1+1}{1}} \\ &= \frac{\binom{2}{1}\binom{2}{2}}{\binom{5}{1}} \\ &= \frac{\frac{2!}{1!\,1!} \times \frac{2!}{2!\,0!}}{\frac{5!}{4!1!}} \\ &= \frac{2}{5} \end{aligned}$$

It is not surprising after all that Jaynes and I have maintained coherency in our answers. First of all, I closely analyzed the essence of Jaynes's proof and followed his direction when developing my own version. Secondly, both he and I took Laplace's advice to heart and made the prior probability of models a uniform probability.

11.5 Coherency as M, N, and n Change

Let's see if things hang together in a coherent fashion as we start to examine more complex inferential scenarios with changing and larger values of M, N, and n. Transition from binary outcomes when tossing coins to the six defined outcomes when rolling a die. The state space has been bumped up from $n = 2$ to $n = 6$. The data happen to consist of four previous rolls of the die with a ONE, two THREEs, and a FIVE appearing. The amount of data has increased from $N = 3$ to $N = 4$. Suppose now that the posterior predictive probability looks ahead to not just the very next roll of the die, but to the next two rolls. M has been bumped up from $M = 1$ to $M = 2$. What is the probability for a SIX and a FOUR to land face up on the next two rolls?

The IP assumes nothing about the physical construction of the "die," in fact, it has never examined or even seen the die that is being rolled. Furthermore, the IP knows nothing about how "rolling a die" has been defined. For all it knows, "rolling a die" means holding the SIX face up 1 mm from a table top and releasing it. Therefore, the IP adopts a state of knowledge about what might be causing any of the six faces to land face up to reflect the fact that it is "totally uninformed."

The IP is aware that under such paucity of knowledge the probability of any face on the *next* throw is 1/6. Consider another IP who, contrary to the one just described, knows a lot about the die and how it will be thrown. The die, in fact, is about as perfect a cube as could be imagined with no physical imperfections whatsoever. The die will be thrown in such a manner that any initial conditions will be completely swamped. As a consequence of this state of knowledge, the IP adopts one single model, the fair model, also with a probability of any face on the *next* throw of 1/6.

Their differing states of knowledge reveal themselves, however, when asked about more than the very next roll of the die. The first IP correctly asserts its probability of a SIX and a FOUR as 1/21. The second IP correctly asserts its probability of a SIX and a FOUR as 1/18. Assume that both IPs labor under the constraint of no observations.

The first IP, although it started out as completely uninformed about the die and its manner of being thrown, is an empiricist of the first rank. It will reason via a generalization of Classical Logic by relying upon the formal manipulation rules of probability. It will update its state of knowledge by taking account of what actually happened during those four previous rolls. It will change its posterior predictive probability from 1/21 when asked about any future outcome.

Curiously, or not, the second IP will not change its prediction. By definition, it already knows with absolute certainty that the die is governed by the fair model. Since it has adopted one single model, no subsequent data can change its degree of belief in that model. The probability for any face remains at 1/6 at every single roll no matter what the data might show.

Supplemental Exercise 11.5.1: What is an IP's degree of belief in the truth of the statements comparing the same face vs. different faces from the perspective of the traditional sample space?

Solution to Supplemental Exercise 11.5.1:

The origin of that single model assignment of 1/6 for any face to appear lies in the orthodox description of the number of elementary points in a sample space. The sample space here consists of $n^M = 6^2 = 36$ elementary points.

These 36 elementary points are the most basic description of what could be observed. One elementary point is: A FOUR occurs on the first roll and a SIX occurs on the second roll. Another elementary point is: A THREE occurs on both rolls.

The total of 36 elementary points can be decomposed into two classes, 1) different faces, 2) same face. There are fifteen ways for the first class to occur; the first elementary point just mentioned is a member of that class. There are six ways for the second class; the second elementary point just mentioned is a member of that class.

For the first class, there are two ways for different faces to occur; for the second class, there is only way. A FOUR and SIX is a member of the first class and it can come about in two ways. The first way is the first elementary point above, and the second way is a SIX on the first roll and a FOUR on the second roll. This case is another elementary point from the total of 36. The total number of elementary points in the sample space is thus accounted for with $36 = (15 \times 2) + (6 \times 1)$ and the probability accounted for with $(15 \times \frac{1}{18}) + (6 \times \frac{1}{36}) = 1$.

The probability for an *event* is defined by simply counting up the number of elementary points that would define that event, and then dividing by the total number of elementary points. This summarizes the situation for the second IP who adopts just one single model, the fair model. The probability of the same face over two rolls of the die is $6/36 = 1/6$. The probability of different faces over two rolls is $30/36 = 5/6$.

For the totally uninformed IP, the probability of the same face over the two rolls is almost twice as probable at 6/21. The probability of different faces is 15/21. There are only $15 + 6 = 21$ possibilities possessing equal probabilities, not 36. The multiplicity factor of 2 for different faces is negated under complete ignorance.

$$P(M_1 = 0, \cdots, M_4 = 1, \cdots, M_6 = 1 \,|\, N_1 = 0, \cdots, N_6 = 0) = \frac{M!\,(n-1)!}{(M+n-1)!} = \frac{2!\,5!}{7!} = \frac{1}{21}$$

Of course, under either perspective, the probability of observing either the same face or different faces in two rolls is a certainty, and the probability must sum to 1.

Supplemental Exercise 11.5.2: Develop at least one coherent argument for the posterior predictive probability as M and N change.

Solution to Supplemental Exercise 11.5.2:

As always, we start out with a conditional probability, the left hand side of Bayes's Theorem,

$$P(A_5 = a_6, A_6 = a_4 \,|\, A_1 = a_3, A_2 = a_1, A_3 = a_5, A_4 = a_3)$$

to express the probability of a SIX and a FOUR on an upcoming fifth and sixth roll of the die. The die has already been rolled four times with a THREE on the first roll, a ONE on the second roll, a FIVE on the third roll, and another THREE on the fourth and final roll.

In a previous exercise where $M = 1$, we discovered that whether the IP conditions on a specific sequence of data as above, or just provides a summary of the number of occurrences, we end up with the same probability. Things that needed to cancel out did cancel out. But that was for the case where $M = 1$. Will it hold when M, the number of future events to be predicted, increases?

The first argument will rely on the familiar posterior predictive formula with,

$$M = 2 \text{ and } M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 1, M_5 = 0, M_6 = 1$$

$$N = 4 \text{ and } N_1 = 1, N_2 = 0, N_3 = 2, N_4 = 0, N_5 = 1, N_6 = 0$$

$$n = 6$$

Filling in the formula for the particulars of this example,

$$P(M_1, \cdots, M_6 \,|\, N_1, \cdots, N_6) = C \times \frac{\prod_{i=1}^{6}(M_i + N_i)!}{\prod_{i=1} M_i!}$$

$$\begin{aligned} C &= \frac{M!\,(N+n-1)!}{N_1! \cdots N_6!\,(M+N+n-1)!} \\ &= \frac{2!\,(4+6-1)!}{1! \cdots 0!\,(2+4+6-1)!} \\ &= \frac{9!}{11!} \end{aligned}$$

$$\frac{\prod_{i=1}^{6}(M_i + N_i)!}{\prod_{i=1} M_i!} = \frac{1!\;0!\;2!\;1!\;1!\;1!}{0!\;0!\;0!\;1!\;0!\;1!}$$

$$C \times \frac{\prod_{i=1}^{6}(M_i + N_i)!}{\prod_{i=1} M_i!} = \frac{2!\;9!}{11!} = \frac{1}{55}$$

Through this argument, we have the posterior predictive probability of a FOUR and SIX on the next two rolls as 1/55 after the already observed data of the first four rolls,

$$P(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 1, M_5 = 0, M_6 = 1 \,|\, \mathcal{D}) = \frac{1}{55}$$

Supplemental Exercise 11.5.3: Develop another argument that must be coherent with the probability as found above.

Solution to Supplemental Exercise 11.5.3:

This argument, contrary to the one just presented in the previous exercise, will condition on the specific sequence of the data. Furthermore, it will go off on a different tack and rely on an argument showing the result as an average over all model predictions as weighted by the posterior probability for the models.

$$P(A_5, A_6 \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_5, A_6 \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k \,|\, \mathcal{D})$$

The argument as it unfolds would not be considered an "elegant" argument under any stretch of the imagination. But, in its defense, no matter what kind of strange gyrations our minds may take in trying to come to grips with understanding a problem, it must ultimately be coherent with whatever elegant argument you do happen to prefer.

The following steps are similar to the ones carried out in Exercise 11.3.1 of these Supplemental Exercises. Use Bayes's Theorem on the posterior probability for the models $P(\mathcal{M}_k \,|\, \mathcal{D})$,

$$\sum_{k=1}^{\mathcal{M}} P(A_5, A_6 \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(A_5, A_6 \,|\, \mathcal{M}_k) \times \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(\mathcal{D})}$$

Bring out the constant factor involving the probability for the data, and transition to a multiple integration over the q_i,

$$\sum_{k=1}^{\mathcal{M}} P(A_5, A_6 \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{1}{P(\mathcal{D})} \times \int \cdot \int_{\mathcal{R}} P(A_5, A_6 \,|\, \mathcal{M}_k)\, P(\mathcal{D} \,|\, \mathcal{M}_k)\, \mathrm{pdf}\,(q_i)\; \mathrm{d}q_i$$

Concentrate now on the integral on the right hand side. Substitute the probabilities under the models for the two future occurrences, that is, of a SIX face followed by a FOUR face. Substitute the probabilities under the models for the four data observations. Substitute the Dirichlet distribution for the prior probability of the statements made by the models, that is, the assertions that the q_i are the correct assignments,

$$\int \cdot \int_{\mathcal{R}} P(A_5, A_6 \,|\, \mathcal{M}_k) P(\mathcal{D} \,|\, \mathcal{M}_k) \,\mathrm{pdf}\,(q_i)\,\mathrm{d}q_i = \frac{1}{P(\mathcal{D})} \times \int \cdot \int_{\mathcal{R}} q_6 \times q_4 \times q_3 \times q_1 \times q_5 \times q_3 \times$$
$$C_D \times q_1^{\alpha_1 - 1} \times q_2^{\alpha_2 - 1} \times q_3^{\alpha_3 - 1} \times$$
$$q_4^{\alpha_4 - 1} \times q_5^{\alpha_5 - 1} \times q_6^{\alpha_6 - 1}\,\mathrm{d}q_i$$

Pull out the constant C_D from under the integral, and add exponents for the q_i,

$$\int \cdot \int_{\mathcal{R}} P(A_5, A_6 \,|\, \mathcal{M}_k) P(\mathcal{D} \,|\, \mathcal{M}_k) \,\mathrm{pdf}\,(q_i)\,\mathrm{d}q_i = \frac{1}{P(\mathcal{D})} \times C_D \times \int \cdot \int_{\mathcal{R}} q_1^{1+\alpha_1 - 1} \times q_2^{\alpha_2 - 1} \times$$
$$q_3^{2+\alpha_3 - 1} \times q_4^{1+\alpha_4 - 1} \times q_5^{1+\alpha_5 - 1} \times q_6^{1+\alpha_6 - 1}\,\mathrm{d}q_i$$
$$= \frac{1}{P(\mathcal{D})} \times C_D \times \int \cdot \int_{\mathcal{R}} q_1 \times q_3^2 \times q_4 \times q_5 \times q_6\,\mathrm{d}q_i$$

There is the known analytical solution to this multiple integral. This is fortunate because otherwise we wouldn't know the Dirichlet distribution,

$$\int \cdot \int_{\mathcal{R}} q_1 \times q_3^2 \times q_4 \times q_5 \times q_6\,\mathrm{d}q_i = \frac{1!\,0!\,2!\,1!\,1!\,1!}{11!}$$

Finally, we can put everything back together and simplify with,

$$P(A_5 = \mathrm{SIX}, A_6 = \mathrm{FOUR} \,|\, \mathcal{D}) = \frac{1}{P(\mathcal{D})} \times C_D \times \frac{2}{11!}$$
$$\frac{1}{P(\mathcal{D})} = \frac{(N+n-1)!}{N!\,(n-1)!}$$
$$C_D = (n-1)!$$
$$P(A_5 = \mathrm{SIX}, A_6 = \mathrm{FOUR} \,|\, \mathcal{D}) = \frac{9!}{4!\,5!} \times 5! \times \frac{2}{11!}$$
$$= \frac{2}{4! \times 11 \times 10} \neq \frac{1}{110}$$

Here is where the argument takes a funny turn and why coherency is so important. When we used the symbol $\mathcal{D}$ for the data, and substituted its probability $P(\mathcal{D})$, we need to account for the fact that this usage of "data" meant the fact that a ONE, two THREEs, and a FIVE occurred. The multiplicity factor $W(N)$ was always present in the derivation for $P(\mathcal{D})$.

But in the above example, the conditioning was on the exact observed temporal sequence of THREE ONE FIVE THREE on the first roll through the fourth roll. The multiplicity factor was missing! Thus,

$$W(N) = \frac{4!}{1!\ 0!\ 2!\ 0!\ 1!\ 0!} = 4 \times 3$$

must appear in the computation because it also appeared in the derivation of $P(\mathcal{D})$. The inclusion of $W(N)$ leads to the cancellation of the 4! term in the denominator,

$$P(A_5 = \text{SIX}, A_6 = \text{FOUR} \,|\, \mathcal{D}) = (4 \times 3) \times \frac{2}{4! \times 11 \times 10} = \frac{1}{110}$$

Recall that when we solved for the future appearance of a SIX and a FOUR on the next two rolls,

$$P(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 1, M_5 = 0, M_6 = 1 \,|\, \mathcal{D}) = \frac{1}{55} = \frac{2}{110}$$

there was a multiplicity factor $W(M)$ present as well. The SIX could occur on the first and the FOUR on the second of future rolls, or the other way around. Each of these two events have a probability of 1/110 from the above result.

Having full faith in the correctness of the posterior predictive probability formula where the multiplicity factors $W(N)$ and $W(M)$ were both present, meant that any coherent argument based on averaging predictions over the posterior model probabilities, the specific order of the prediction, when given the faces on specific rolls as data, must have a probability of 1/110.

When $M > 1$, and the exact sequence of events is spelled out, then care must be paid to what is happening with the multiplicity factors $W(N)$ and $W(M)$ in any argument.

11.6 Causal Factors in Coin Tossing

We return to the coin tossing scenario for the next two exercises concerning causal factors. In section 11.4 of Volume I, an amendment was introduced to the posterior predictive formula to cover the situation when a causal factor was present.

The symbolic expression for the conditional probability of HEADS on the next toss given that N tosses have already been made, and some causal factor B is present and readily observable at trial $N + 1$, looks like this,

$$P(A_{N+1} = \text{HEADS} \,|\, B_{N+1} = \text{Present}, \mathcal{D})$$

This modification to the formula was mandated solely by a rigorous adherence to the formal manipulation rules.

The state space must be enlarged from the original $n = 2$ to $n = 4$ because there are now four possible measurements at each toss.

1. $(A = a_1)$ and $(B = b_1)$, "HEADS and causal factor present,"

2. $(A = a_2)$ and $(B = b_1)$, "TAILS and causal factor present,"

3. $(A = a_1)$ and $(B = b_2)$, "HEADS and causal factor absent," and finally

4. $(A = a_2)$ and $(B = b_2)$, "TAILS and causal factor absent."

Supplemental Exercise 11.6.1: In the interests of a concrete numerical example, assume that an experiment consisting of $N = 10,000$ coin tosses has been conducted.

Solution to Supplemental Exercise 11.6.1:

The contingency table, shown below in Figure 11.1, is constructed with four cells and illustrates the data collected over the $N = 10,000$ trials in the experiment. Each cell contains the N_i frequency counts recorded for the joint statement of face showing and status of the causal factor.

	HEADS A	TAILS $\overline{A}$	
CF Present B	4000 N_1	1000 N_2	5000
CF Absent $\overline{B}$	2500 N_3	2500 N_4	5000
	6500	3500	10000

Figure 11.1: *Contingency table for 10,000 fictitious coin tosses.*

It looks as if the causal factor has something to do with increasing the frequency of HEADS. Suppose, for the sake of this example, the causal factor is identified as a person who has a special skill in making HEADS appear in coin tosses. When this person is not tossing the coin, the causal factor is absent, and the coin reverts to a "fair coin."

I take it you are not particularly surprised when we start off with an application of Bayes's Theorem,

$$P(A_{N+1} = \text{HEADS} \,|\, B_{N+1} = \text{Present}, \mathcal{D}) = \frac{P(A_{N+1}, B_{N+1} \,|\, \mathcal{D})}{P(B_{N+1} \,|\, \mathcal{D})}$$

$$= \frac{P(A_{N+1}, B_{N+1} \,|\, \mathcal{D})}{P(A_{N+1}, B_{N+1} \,|\, \mathcal{D}) + P(\overline{A}_{N+1}, B_{N+1} \,|\, \mathcal{D})}$$

Laplace's **Rule of Succession** will allow us to quite easily fill in these terms in the numerator and denominator of Bayes's Theorem,

$$P(A_{N+1}, B_{N+1} \,|\, \mathcal{D}) = \frac{N_1 + 1}{N + n}$$

$$= \frac{4001}{10004}$$

$$P(\overline{A}_{N+1}, B_{N+1} \,|\, \mathcal{D}) = \frac{N_2 + 1}{N + n}$$

$$= \frac{1001}{10004}$$

$$P(A_{N+1} = \text{HEADS} \,|\, B_{N+1} = \text{Present}, \mathcal{D}) = \frac{4001/10004}{(4001/10004) + (1001/10004)}$$

$$= \frac{4001}{5002}$$

If there is any justification for thinking about probabilities in terms of frequencies, it is here. Jaynes's remark [11, pg. 571] is apropos:

> In the literature starting with Venn (1866), those who issued polemical denunciations of Laplace's rule of succession have put themselves in an incredible situation. How is it possible for one human mind to reject Laplace's rule—and then advocate a frequency definition of probability? Anyone who assigns a probability to an event equal to its observed frequency in many trials is doing just what Laplace's rule tells him to do! The generalized rule [my general posterior prediction formula] supplies an obviously needed refinement of this, small correction terms when the number of observations is not large compared with the number of propositions. [When N is not large compared to n.]

Supplemental Exercise 11.6.2: Redo the above exercise more formally by spelling out all of the details in the posterior prediction formula.

Solution to Supplemental Exercise 11.6.2:

For this problem we have the following specifications,

$$n = 4, M = 1, N = 10,000, N_1 = 4000, N_2 = 1000, N_3 = 2500, N_4 = 2500$$

Any M_i also has to equal 1, and the particular M_i in question is determined by the future joint statement $(A = a_i, B = b_j)$ in question.

Examine the numerator in Bayes's Theorem,

$$P(A_{N+1}, B_{N+1} \,|\, \mathcal{D}) \equiv P(A_{10001} = \text{HEADS}, B_{10001} = \text{Present} \,|\, N_1, N_2, N_3, N_4)$$

The way the contingency table was constructed, this joint statement occurs in cell 1. Thus, $M_1 = 1$, $M_2 = 0$, $M_3 = 0$, and $M_4 = 0$.

$$\begin{aligned}
P(A_{N+1}, B_{N+1} \,|\, \mathcal{D}) &= P(M_1 = 1, M_2 = 0, M_3 = 0, M_4 = 0 \,| \\
&\qquad\qquad N_1 = 4000, N_2 = 1000, N_3 = 2500, N_4 = 2500) \\
&= \frac{M!\,(N+n-1)!}{N_1!\,N_2!\,N_3!\,N_4!\,(M+N+n-1)!} \times \frac{\prod_{i=1}^{4}(M_i+N_i)!}{\prod_{i=1}^{4} M_i!} \\
&= \frac{1!\,(10000+4-1)!}{4000!\,1000!\,2500!\,2500!\,(1+10000+4-1)!} \times \frac{4001!\,1000!\,2500!\,2500!}{1!\,0!\,0!\,0!} \\
&= \frac{10003!}{4000!\,1000!\,2500!\,2500!\,10004!} \times 4001!\,1000!\,2500!\,2500! \\
&= \frac{4001}{10004}
\end{aligned}$$

And, of course, the probabilities for the future occurrence of any of the other three joint statements would be calculated in exactly the same way. The particular $M_i = 1$ would change depending upon the future joint statement.

Chapter 12

Extending the Formal Rules for Prediction

12.1 Conceptual Distinctions

My derivation of the posterior predictive probability for future frequency counts as first presented in Chapter Twelve in the concise formula,

$$P(M_1, M_2, \cdots, M_n \,|\, \mathcal{D}) = C \times \frac{\prod_{i=1}^{n}(M_i + N_i)!}{\prod_{i=1}^{n} M_i!}$$

[12] 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.

was inspired by closely examining three sources.

The first source was Jaynes, mainly in two expositions. The earlier one to appear in print chronologically, although I'm not sure it was prior in the evolution of his thoughts on this matter, was the MKN paper [12]. Jaynes's second treatment of the posterior predictive probability appeared in Chapter 18 of his book [11]. I record my debt to Jaynes in helping me sort out the posterior predictive probability in various places throughout my books.

The second source was Seymour Geisser's small introductory book on predictive inference [9]. This book was valuable to me mainly because of his level of detail in examples where he used a *beta distribution* as a prior probability over model space. Deconstructing Geisser's derivations for prior and posterior predictive distributions confirmed my tentative forays in the same area. They also were in concordance with what I had gleaned previously from Jaynes's two works just mentioned.

The third source for refining my initially confused and ill–formed suppositions about the posterior predictive probability came from the introductory section of Bernardo & Smith's Chapter 5 [1, pp. 241–244].

In fact, I even borrowed Bernardo & Smith's notation in my Exercises 12.6.11 through 12.6.13 of Volume I as a superior concise introduction to the underlying

probability manipulations taking place in the derivation. However, in this case, I was not quite as effusive when acknowledging their contributions to my cause. What was the reason for this omission?

Even though it may seem, at first blush, that the foundation supporting all of the probability manipulations seem to be identical, there arose glaring conceptual anomalies that I could not paper over. I believe it is helpful to *closely* deconstruct the *actual language relevant to fundamental concepts* used by Bernardo & Smith and compare it to my alternative language about concepts.

Supplemental Exercise 12.1.1: Quote Bernardo & Smith on how they would like to begin the discussion of the "Bayesian paradigm."

Solution to Supplemental Exercise 12.1.1:

On page 241 [1], they begin with,

> **5.1 THE BAYESIAN PARADIGM**
>
> **5.1.1 Observables, Beliefs and Models**
>
> Our development has focused on the foundational issues which arise when we aspire to formal quantitative coherence in the context of decision making in situations of uncertainty. This development, in combination with an operational approach to the basic concepts, has led us to view the problem of statistical modeling as that of identifying or selecting particular forms of representation of beliefs about observables.

Now, I would not quibble with this introduction, and I think it would be fair to say that these thoughts reflect my intentions as well. Unfortunately, after this auspicious beginning, we immediately diverge in coming to grips with supposed complications any initial approach must face up to.

Supplemental Exercise 12.1.2: What is the very first technical notion Bernardo and Smith trot out after the above introduction?

Solution to Supplemental Exercise 12.1.2:

At the top of page 242, they present us with this notion,

> For example, in the case of a sequence $x_1, x_2, \ldots$, of 0–1 random quantities for which beliefs correspond to a judgment of infinite exchangeability, Proposition 4.1, (de Finetti's theorem) identifies the representation of the joint mass function for $x_1, \ldots, x_n$ as having the form
>
> $$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^n \theta^{x_i} \, (1-\theta)^{1-x_i} \, dQ(\theta)$$
>
> for some choice of distribution Q over the interval [0, 1].

> ...
>
> Such representations, and the more complicated forms ... exhibit the various ways in which the element of primary significance from the subjectivist, operationalist standpoint, namely the *predictive model* of beliefs about observables, can be thought of *as if* constructed from a *parametric model* together with a *prior distribution* for the labeling parameter.
>
> Our primary concern in this chapter will be with the way the updating of beliefs in the light of new information takes place within the framework of each representation. [Emphasis in the original.]

I ask you: Are we really forced to delve into the obscure pedantry of language like "0–1 random variables," "infinite exchangeability," and "de Finetti's representation theorem" so early on in the game?

I can arrive at the same expression just by adhering to our non–mysterious definitions and formal manipulation rules as they have been developed thus far. The following discursive language will be quite familiar, repetitive, and I hope even somewhat boring to those who have travelled along with me on the journey to date.

Immediately dispense with the language of "0–1 random variables" whatever those might be. Inferential problems always center on *statements* existing in a state space of dimension n. Here, for "0–1 random variables," the state space is only of dimension $n = 2$. The coin flip scenario consisting of only two possible statements is as good as any, and more familiar than most.

The expression $(X_t = x_2)$ is defined as the statement, "At the tth trial, or repetition, or measurement, or observation, the coin showed TAILS." In words, the probability expression $P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1)$ is interpreted as: **The degree of belief as held by an IP that the following joint statement is TRUE, "HEADS occurs on the first trial, ..., and TAILS occurs on the tth trial, ..., and HEADS occurs on the Nth and final trial."**

Recall the following series of steps that advance us from this primary expression to an expression equivalent to Bernardo and Smith's. My personal criterion is that every step must be immediately obvious, non–mysterious, and must have already been introduced as part of the axiomatic conceptual armamentarium available to probability theory.

The probability expression that we began with, namely,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1)$$

is an *abstract* probability indicating that the information which would provide a numerical assignment through some model has been hidden away. We unpack this implicit hidden information through the **Sum Rule** together with a familiar axiom borrowed from Boolean Algebra, the ubiquitous **Commutativity axiom**,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1) = \sum_{k=1}^{\mathcal{M}} P(X_N, X_{N-1}, \cdots, X_t, \cdots, X_1, \mathcal{M}_k)$$

I now commence with that hopefully clear series of steps that takes us to the "de Finetti representation theorem." Invoke the **Product Rule** to turn the above into,

$$\begin{aligned}\sum_{k=1}^{\mathcal{M}} P(X_N, X_{N-1}, \cdots, X_t, \cdots, X_1, \mathcal{M}_k) = \sum_{k=1}^{\mathcal{M}} P(X_N \,|\, X_{N-1}, \cdots, X_1, \mathcal{M}_k) \quad &\times \\ P(X_{N-1} \,|\, X_{N-2}, \cdots, X_1, \mathcal{M}_k) \quad &\times \\ \cdots \quad &\times \\ P(X_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \quad &\end{aligned}$$

The very next step is conceptually critical. I have a strong feeling that the confused notion of "exchangeability" is invoked to mimic what is a very clear notion. The probability of an occurrence at any trial depends solely on how the state space was originally defined, together with the information in the model, but ignoring what happened on any number of previous trials.

Do you really think that HEADS showing up on the tenth trial is causally impacted by the fact that TAILS showed up on the ninth trial? If one adopts the view that flipping a coin is no more and no less than a manifestation of ordinary mechanics and physics, then the causal factors must be any relevant physical factor like shape, weight, symmetry, center of gravity, initial orientation, momentum, and so on.

The fact that TAILS showed up on the ninth trial indicates that these physical factors were such that they caused TAILS to appear. The physical factors obtaining on the tenth trial may be completely different, or completely the same, or somewhat the same and somewhat different, as the ninth trial; we simply don't know. But it is certain that mere fact of TAILS appearing on the ninth trial will NOT cause HEADS to appear on the tenth trial.

Therefore, this argument that X_t is independent of any previous $X_{t-1}, \cdots, X_1$ leads us to consider the simplification where only the information in the given model $\mathcal{M}_k$ can influence the assigned probability at any trial. This permits us to simplify the above step to,

$$\begin{aligned}\sum_{k=1}^{\mathcal{M}} P(X_N, X_{N-1}, \cdots, X_t, \cdots, X_1, \mathcal{M}_k) = \sum_{k=1}^{\mathcal{M}} P(X_N \,|\, \mathcal{M}_k) \times P(X_{N-1} \,|\, \mathcal{M}_k) \quad &\times \\ \cdots \quad &\times \\ P(X_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \quad &\end{aligned}$$

At the next step, as all of Volume II tried to explain, the assigned numerical probability to any of the above terms, $P(X_t = x_i \,|\, \mathcal{M}_k)$, is given an operational meaning through the Maximum Entropy Principle.

Repeating the critical foundational concept: **The numerical value of the probability is assigned by the information resident in the model $\boldsymbol{\mathcal{M}_k}$, and not by the previous occurrences of $\boldsymbol{X_t}$.**

Adopting our usual notation with the MEP formula on the right,

$$Q_1 \equiv P(X = x_1 \,|\, \mathcal{M}_k) = \frac{e^{\lambda F(X=x_1)}}{e^{\lambda F(X=x_1)} + e^{\lambda F(X=x_2)}}$$

$$Q_2 \equiv P(X = x_2 \,|\, \mathcal{M}_k) = \frac{e^{\lambda F(X=x_2)}}{e^{\lambda F(X=x_1)} + e^{\lambda F(X=x_2)}}$$

With $n = 2$, the next step looks like this,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1) = \sum_{k=1}^{\mathcal{M}} Q_1 \times \cdots \times Q_2 \times \cdots \times Q_1 \times P(\mathcal{M}_k)$$

$$= \sum_{k=1}^{\mathcal{M}} Q_1^{N_1} \times Q_2^{N_2} \times P(\mathcal{M}_k)$$

This indicates that HEADS occurred N_1 times and TAILS occurred N_2 times where the total number of trials was $N = N_1 + N_2$.

Now, Q_1 and Q_2 are going to change assignment values with every change in the model $\mathcal{M}_k$. By definition, we are summing over all models where an assignment could range from, thinking first discretely before transitioning to the continuous, $Q_1 = 1$ and $Q_2 = 0$ under the first model, $Q_1 = 0.999$ and $Q_2 = 0.001$ under the second model, all the way through to $Q_1 = 0$ and $Q_2 = 1$ under the final model.

Making that transition to a continuous model space, we switch from a summation to an integration,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1) = \int_0^1 q^{N_1} (1-q)^{N_2} \; P(\mathcal{M}_k) \; \mathrm{d}q$$

I didn't completely switch over to a continuous expression because I wanted to focus on the prior probability over model space $P(\mathcal{M}_k)$, and left it as it last appeared in its discrete form.

Analytically, the integration is solvable when the prior probability is in what is called the *conjugate* form to the probability for the sequence of HEADS and TAILS appearing first in the integrand. The conjugate form for a prior probability over model space is the *beta distribution*. When this probability density function pdf (q) is substituted for $P(\mathcal{M}_k)$, we are left with,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1) = \int_0^1 q^{N_1} (1-q)^{N_2} \; C_{\text{Beta}} \; q^{\alpha-1} \, (1-q)^{\beta-1} \; \mathrm{d}q$$

This is exactly what we wanted at the conclusion of our derivation because, despite the notational differences, we are at the same place as Bernardo and Smith's probability for a sequence of "0–1 random variables,"

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} \, (1-\theta)^{1-x_i} \, dQ(\theta)$$

where it all started.

The important lesson from this alternative derivation is not a subtle one. There was no need for beliefs in judgments of infinite exchangeability, nor any appeal to de Finetti's theorem, random quantities, or any other of the dubious notions dragged in to satisfy their argument.

All that was required was a strict adherence to the rules of the game as laid down for any inferential problem. These rules of the game are what have been slowly developed over the course of these books. If an IP can't solve a problem like this after all of the hard work in generalizing Boolean Algebra and Classical Logic as encapsulated within the formal manipulation rules of probability, then what was the effort worth?

I finish up my commentary in this exercise by returning to, and then critiquing, Bernardo & Smith's characterization of the Bayesian paradigm when they employ language like this: "... the predictive model of beliefs about observables can be thought of *as if* constructed from a parametric model together with a prior distribution for the labeling parameter."

I certainly do agree with the first part that an IP would indeed like to justify its beliefs about observables partly from a parametric model. But I disagree about the complementary partner in the enterprise. The *prior distribution* is a prior distribution over all conceivable models, not a prior distribution over a labeling parameter.

In order to emphasize this distinction, closely scrutinize the symbolic expressions for the prior probability. For Bernardo & Smith, it is $p(\boldsymbol{\theta})$. Whereas, for me, it is instead $P(\mathcal{M}_k) \equiv \text{pdf}\,(q_1, q_2, \cdots, q_n)$. Bernardo & Smith's "labeling parameters" $\boldsymbol{\theta}$ are my numerical assignments to statements conditioned on assuming the truth of the information inserted by some model. Thus,

$$\theta \equiv Q_1 = P(X = x_1 \,|\, \mathcal{M}_k) \text{ and } (1-\theta) \equiv Q_2 = P(X = x_2 \,|\, \mathcal{M}_k)$$

These conditional probabilities are obviously **NOT** equivalent to the parameters from the MEP, the Lagrange multipliers λ_j. A prior probability is **NEVER** placed on the MEP's parameters.

Supplemental Exercise 12.1.3: Finish up by finding the computational formula for the probability of the sequence of occurrences of HEADS and TAILS over N tosses.

Solution to Supplemental Exercise 12.1.3:

A few more transformations on the formula for the probability of any sequence of N_1 HEADS and N_2 TAILS over a total of N flips of the coin, yields this computational formula. In the last exercise, we were left with,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1) = \int_0^1 q^{N_1}(1-q)^{N_2}\, C_{\text{Beta}}\, q^{\alpha-1}\,(1-q)^{\beta-1}\, \mathrm{d}q$$

Working on the right hand side, extract the constant C_{Beta} from underneath the integral sign, and then add the exponents for q and $(1-q)$,

$$\int_0^1 q^{N_1}(1-q)^{N_2}\, C_{\text{Beta}}\, q^{\alpha-1}\,(1-q)^{\beta-1}\, \mathrm{d}q = C_{\text{Beta}} \times \int_0^1 q^{N_1+\alpha-1}(1-q)^{N_2+\beta-1}\, \mathrm{d}q$$

The solution for the beta integral is known,

$$C_{\text{Beta}} \times \int_0^1 q^{N_1+\alpha-1}(1-q)^{N_2+\beta-1}\, \mathrm{d}q = C_{\text{Beta}} \times \frac{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)}{\Gamma(N_1+N_2+\alpha+\beta)}$$

Similarly, substituting for the constant term C_{Beta} and $N = N_1 + N_2$, results in an easily implementable computational formula,

$$C_{\text{Beta}} \times \frac{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)}{\Gamma(N_1+N_2+\alpha+\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \times \frac{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)}{\Gamma(N+\alpha+\beta)}$$

We can now compute the probability for any sequence of binary statements over N observations as,

$$P(X_1 = x_1, \cdots, X_t = x_2, \cdots, X_N = x_1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \times \frac{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)}{\Gamma(N+\alpha+\beta)} \tag{12.1}$$

Supplemental Exercise 12.1.4: What conceptual distinction must be kept front and center when contemplating the last few exercises?

Solution to Supplemental Exercise 12.1.4:

The clue is provided in the last exercise where N, the total number of *past* frequency counts, together with N_1 and N_2, were employed instead of M, the total number of *future* frequency counts, together with M_1 and M_2. Equation (12.1) should be explicitly identified as the probability for already observed data in a specified sequence. This probability is conceptually different than $P(M_1, M_2)$ even though the expressions are exactly the same after switching N, N_1, N_2 for M, M_1, M_2.

Supplemental Exercise 12.1.5: Furnish an easy example of the above conceptual distinction that very many people, even those schooled in probability and statistics, find shocking and scarcely believable.

Solution to Supplemental Exercise 12.1.5:

If an IP is totally uninformed about the physical nature of a coin and the manner in which it will be flipped, then it will think it twice as likely that both flips will result in the same face appearing as opposed to different faces.

This assertion is absolutely true if you believe in probability theory. It is not an opinion. It is not some whim issuing from my personal preference. There is no, "Well, if you look at it from a different perspective ... " The result cannot be denigrated by name–calling; is not some "Bayesian subjectivist interpretation," or any other kind of qualifying caveat. No other answer is justifiable.

Consider the probabilities for any sequence of binary statements. Here, we want the probability for a future sequence of outcomes for a coin flip where, by consensual agreement, the state space has been restricted to dimension $n = 2$. The only possible observations for an outcome of a coin flip are HEADS or TAILS.

Even for so simple of a scenario where the coin will be flipped just twice in the future, that is, $M = 2$, the results are eye–opening. The left hand side expression for the joint probability of any sequence becomes for, say, TAILS on the first flip and HEADS on the second,

$$P(X_1, \cdots, X_t, \cdots, X_M) \equiv P(X_1 = \text{TAILS}, X_2 = \text{HEADS})$$

where $M_1 = 1$ and $M_2 = 1$. Making use of our formula, again with the switch to M, M_1, and M_2 evident,

$$\begin{aligned}
P(X_1 = \text{TAILS}, X_2 = \text{HEADS}) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \times \frac{\Gamma(M_1+\alpha)\,\Gamma(M_2+\beta)}{\Gamma(M+\alpha+\beta)} \\
&= \frac{\Gamma(1+1)}{\Gamma(1)\,\Gamma(1)} \times \frac{\Gamma(1+1)\,\Gamma(1+1)}{\Gamma(2+1+1)} \\
&= \frac{\Gamma(2)}{\Gamma(1)\,\Gamma(1)} \times \frac{\Gamma(2)\,\Gamma(2)}{\Gamma(4)} \\
&= \frac{1}{6}
\end{aligned}$$

In exactly the same manner, the probability for two TAILS to appear is,

$$\begin{aligned}
P(X_1 = \text{TAILS}, X_2 = \text{TAILS}) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \times \frac{\Gamma(M_1+\alpha)\,\Gamma(M_2+\beta)}{\Gamma(M+\alpha+\beta)} \\
&= \frac{\Gamma(1+1)}{\Gamma(1)\,\Gamma(1)} \times \frac{\Gamma(0+1)\,\Gamma(2+1)}{\Gamma(2+1+1)} \\
&= \frac{\Gamma(2)}{\Gamma(1)\,\Gamma(1)} \times \frac{\Gamma(1)\,\Gamma(3)}{\Gamma(4)} \\
&= \frac{1}{3}
\end{aligned}$$

After this, we see that the probability for the other two possibilities are,

$$P(X_1 = \text{HEADS}, X_2 = \text{HEADS}) = \frac{1}{3} \text{ and } P(X_1 = \text{HEADS}, X_2 = \text{TAILS}) = \frac{1}{6}$$

Thus our claim is upheld. The probability of two HEADS or two TAILS is twice as probable as HEADS and then TAILS, or TAILS and then HEADS.

Of course, if a detailed examination were to take place for how this could have happened, there will be no quibbling over M_1, M_2, or M. This kind of *post mortem* will serve to focus the attention on the critical role of the prior probability over model space, $P(\mathcal{M}_k) \equiv \text{pdf}\ (q)$.

There can be only one definition of what "an uninformed IP" means. It is Laplace's definition for an IP who is totally ignorant of the causes that make HEADS or TAILS to appear. It must perforce allocate equal probability to every conceivable "bias" of the coin.

Operationally, this is a probability density function pdf (q) that doesn't favor any assignment for q and $(1-q)$ over any other. This goal is achieved through the invocation of a *beta probability density function* over q where its two parameters α and β are both set equal to 1. The density function is "flat" taking on the functional value of 1 over the entire permissible interval of q from 0 to 1.

Supplemental Exercise 12.1.6: Write some *Mathematica* code to first double–check, and then implement the computational formula.

Solution to Supplemental Exercise 12.1.6:

Have *Mathematica* perform a direct symbolic integration of the expression for the probability of future frequency counts of binary variables over M trials,

```
Integrate[q^M1 (1-q)^M2 PDF[BetaDistribution[α, β], q],
      {q, 0, 1}, Assumptions → α + Re[M1] > 0 && β + Re[M2] > 0]
```

The expression returned by this integration can then be folded in to define a function **f[ ]** with four arguments,

```
f[M1_, M2_, α_, β_] := (Gamma[M1 + α] Gamma[M2 + β]) /
                       (Beta[α, β] Gamma[M1 + M2 + α + β ])
```

The probability for a future sequence of two TAILS over two trials when an IP's state of knowledge about the causes could be characterized as "uninformed" was computed as 1/3. Specifying the four arguments as $M_1 = 0$, $M_2 = 2$, $\alpha = \beta = 1$, this function **f[0, 2, 1, 1]** confirms the probability of 1/3.

Comparing my formula in Exercise 12.1.3 with the *Mathematica* answer, it is obvious that **Beta[α, β]**, the Euler beta function, appears as the term,

$$\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Supplemental Exercise 12.1.7: What is the astonishing conclusion from this formula as more and more trials are considered?

Solution to Supplemental Exercise 12.1.7:

Contrary to the initial general impression you get by reading Jaynes, the multiplicity factor may be totally irrelevant in determining probabilities. Consider first the situation where $M = 9$ flips of the coin will take place.

There is only *one* way for nine successive HEADS to occur, obviously a HEADS must occur on every single flip of the coin. There are 126 ways for five HEADS and four TAILS to occur over the nine coin tosses. The probability of all HEADS when the IP is uninformed about the causes is 1/10. The probability for five HEADS is 1/10. The probability for any number of HEADS from zero to nine is 1/10.

Since there are, in fact, 126 ways to obtain five HEADS and four TAILS, each individual possible sequence from the total of 126, as in HTTHHTHTH, must have the probability of 1/1260 if the compound event of five HEADS and four TAILS is to have a probability of 1/10. Then, the sequence of all nine HEADS is far more probable than any one of the 126 individual sequences leading to five HEADS.

Does the formula confirm the probability for a particular sequence of five HEADS and four TAILS, say, for the sequence given above as an example?

$$\begin{aligned} P(X_1 = \text{HEADS}, X_2 = \text{TAILS}, \cdots, X_9 = \text{HEADS}) &= \frac{\Gamma(2)}{\Gamma(1)\,\Gamma(1)} \times \frac{\Gamma(5+1)\,\Gamma(4+1)}{\Gamma(9+1+1)} \\ &= \frac{5!\;4!}{10!} \\ &= \frac{1}{1260} \end{aligned}$$

Or, evaluating **f[5, 4, 1, 1]** returns the same probability.

Supplemental Exercise 12.1.8: What did Bayes and Laplace have to say on the matter of a prior probability for models?

Solution to Supplemental Exercise 12.1.8:

The Reverend Bayes, in his **Scholium**, said this:

> And that the same rule is the proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials concerning it, seems to appear from the following consideration; *viz.* that concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another.

And the Marquis de Laplace, speaking in French of course, said this:

> Lorsqu'on n'acune donnée *a priori* sur la possibilité d'un événement, il faut supposer toutes les possibilitiés, depuis zéro jusqu'à l'unité, également probables . . . [My very free translation in the context and language of everything we have studied so far: If an IP possesses a state of knowledge about the causes of events such that it knows nothing about those causes, then all conceivable assigned probabilities for the event, that is, all values for q between 0 and 1 must be considered on an equal basis such that no one assigned value for q is favored over any other.]

12.2 Advancing the Bayesian paradigm

Pick up the thread of Bernardo & Smith's development of the "Bayesian paradigm" where we left off before the diversions of the last section. Their straightforward development of the posterior predictive formula is welcome at this point.

As I have already mentioned, studying this portion, and the easily understood invocation of Bayes's Theorem for the equations, assisted me greatly. The only qualifications are the important conceptual qualifications surrounding the meaning of their "parameters $\boldsymbol{\theta}$" together with the "prior probability over parameters" $p(\boldsymbol{\theta})$.

> **5.1.2 The Role of Bayes' Theorem**
> In its simplest form, within the formal framework of predictive model belief distributions derived from quantitative coherence considerations, the problem corresponds to identifying the joint conditional density of
>
> $$p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n)$$
>
> for any $m \geq 1$, given, for any $n \geq 1$, the form of representation of the joint density $p(x_1, \cdots, x_n)$.
>
> In general, of course, this simply reduces to calculating
>
> $$p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n) = \frac{p(x_1, \cdots, x_{n+m})}{p(x_1, \cdots, x_n)}$$

and, in the absence of further structure, there is little more that can be said. However, when the predictive model admits a representation in terms of parametric models and prior distributions, the learning process can be essentially identified, in conventional terminology, with the standard parametric form of Bayes' theorem.

Thus, for example, if we consider the general parametric form of representation for an exchangeable sequence, with $dQ(\boldsymbol{\theta})$ having density representation, $p(\boldsymbol{\theta})d\boldsymbol{\theta}$, we have

$$p(x_1, \cdots, x_n) = \int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

from which it follows that

$$\begin{aligned} p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n) &= \frac{\int \prod_{i=1}^{n+m} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &= \int \prod_{i=n+1}^{n+m} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, x_1, \cdots, x_n)\, d\boldsymbol{\theta} \end{aligned}$$

Supplemental Exercise 12.2.1: Translate Bernardo & Smith's formulas into our notation.

Solution to Supplemental Exercise 12.2.1:

With the expression for a joint conditional density of,

$$p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n)$$

Bernardo & Smith are asking for the probability of a future sequence of binary variables given some already observed data. Suppose that an IP wants to calculate the probability for the sequence of TAILS, HEADS, TAILS in the next three coin tosses *after* the coin had already been tossed nine times previously. The exact sequence for the data will be the same as in Exercise 12.1.6 with five HEADS and four TAILS.

In our notation then, the past data has $N = 9$ with $N_1 = 5$ and $N_2 = 4$ and the future occurrences $M = 3$ with $M_1 = 1$ and $M_2 = 2$. An application of Bayes's Theorem analogous to Bernardo & Smith's joint conditional density results in,

$$P(X_{N+1}, \cdots, X_{N+M} \,|\, X_1, \cdots X_N) = \frac{P(X_1, \cdots, X_{N+M})}{P(X_1, \cdots X_N)}$$

which we rearrange and condense to,

$$P(X_{N+M}, \cdots, X_{N+1} \,|\, \mathcal{D}) = \frac{P(X_{N+M}, \cdots, X_1)}{P(\mathcal{D})}$$

The joint conditional density on the left hand side is specifically,

$$P(X_{12} = \text{TAILS}, X_{11} = \text{HEADS}, X_{10} = \text{TAILS} \,|\, \mathcal{D})$$

The numerator of Bayes's Theorem on the right hand side becomes,

$$\begin{aligned}\int (1-q)\times q\times(1-q)\times q^5\times(1-q)^4\times \text{pdf}\,(q)\,\mathrm{d}q &= \int q^6(1-q)^6\,C_{\text{Beta}}\,q^{\alpha-1}(1-q)^{\beta-1}\,\mathrm{d}q\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\times\frac{\Gamma(6+\alpha)\,\Gamma(6+\beta)}{\Gamma(6+6+\alpha+\beta)}\\ &= 1\times\frac{\Gamma(6+1)\,\Gamma(6+1)}{\Gamma(12+1+1)}\\ &= \frac{6!\;6!}{13!}\end{aligned}$$

The denominator in Bayes's Theorem was discussed in Exercise 12.1.7 with a result that $P(\mathcal{D}) = 1/1260$. Therefore,

$$P(X_{12}=\text{TAILS}, X_{11}=\text{HEADS}, X_{10}=\text{TAILS}\,|\,\mathcal{D}) = 1260\times\frac{6!\;6!}{13!} = 0.104895$$

It's not hard to express the numerator for the sequence of binary statements in general as,

$$\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}\times\frac{\Gamma(N_1+M_1+\alpha)\,\Gamma(N_2+M_2+\beta)}{\Gamma(N+M+2)}$$

My preferred notation for the posterior sequence of statements,

$$P(X_{N+M},\cdots,X_{N+1}\,|\,\mathcal{D}) = \frac{P(X_{N+M},\cdots,X_1)}{P(\mathcal{D})}$$

and my solution just presented is directly analogous to what Bernardo & Smith wrote as,

$$p(x_{n+1},\cdots,x_{n+m}\,|\,x_1,\cdots,x_n) = \frac{\int\prod_{i=1}^{n+m}p(x_i\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int\prod_{i=1}^{n}p(x_i\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

Once again, while there is no disagreement between us over the form that Bayes's Theorem must assume with whatever symbolic probability expressions that appear, there most certainly exists a conceptual divide when the label "parameter" enters into the meaning of the prior probability $p(\boldsymbol{\theta})$ *versus* $P(\mathcal{M}_k)$. For me, the prior probability must be a probability density function over q, for example, the *beta distribution*, and NOT some distribution over the parameter λ appearing in the parametric modeling,

$$Q_i \equiv P(X=x_i\,|\,\mathcal{M}_k) = \frac{e^{\lambda\,F(X=x_i)}}{Z(\lambda)}$$

Supplemental Exercise 12.2.2: What steps did Bernardo & Smith skip over?

Solution to Supplemental Exercise 12.2.2:

In going from,

$$p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n)$$

to,

$$\int \prod_{i=n+1}^{n+m} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, x_1, \cdots, x_n)\, d\boldsymbol{\theta}$$

they skipped over an explicit mention of another application of Bayes's Theorem,

$$p(\boldsymbol{\theta} \,|\, x_1, x_2, \cdots, x_n) = \frac{p(x_1, x_2, \cdots, x_n, \boldsymbol{\theta})}{p(x_1, x_2, \cdots, x_n)} = \frac{p(x_1, x_2, \cdots, x_n \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(x_1, x_2, \cdots, x_n \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

After applying their simplification to both numerator and denominator, which, by the way, is always justified by appeal to "independent and identically distributed variables," they have,

$$p(\boldsymbol{\theta} \,|\, x_1, x_2, \cdots, x_n) = \frac{\prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

The next step left unmentioned was to conveniently parcel out the $n+m$ products in the numerator into two parts, the first from $i = n+1$ to $n+m$, and the second from $i = 1$ to n,

$$p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n) = \frac{\int \prod_{i=n+1}^{n+m} p(x_i \,|\, \boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

We can now carve out the expression for the posterior probability just found above, and leverage it to arrive at the correct formal expression,

$$p(x_{n+1}, \cdots, x_{n+m} \,|\, x_1, \cdots, x_n) = \int \prod_{i=n+1}^{n+m} p(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, x_1, \cdots, x_n)\, d\boldsymbol{\theta}$$

In words, this expression on the right hand side is their way of saying that the posterior predictive probability is an average of the future occurrences with respect to the posterior probability over *parameter* space. My conceptual disagreement is that it should say that the posterior predictive probability is an average of the future occurrences with respect to the posterior probability over *model* space.

My version of the posterior predictive probability would be written as,

$$P(X_{N+1}, \cdots, X_{N+M} \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(X_{N+1}, \cdots, X_{N+M} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, \mathcal{D})$$

Supplemental Exercise 12.2.3: Comment on the ubiquitous appearance of the phrase, "independent and identically distributed variables."

Solution to Supplemental Exercise 12.2.3:

I want to draw attention to the fact that our framework does not automatically jump to the conclusion of "independent and identically distributed variables." Although we have, just like Bernardo & Smith, written out probability expressions like,

$$P(X_1, X_2, \cdots, X_{N+M} \,|\, \mathcal{M}_k) = q^{N_1}\,(1-q)^{N_2}\,q^{M_1}\,(1-q)^{M_2}$$

for a state space of dimension $n = 2$, we do have the option open to us of redefining the state space.

We could always redefine the state space to have dimension 2^{N+M} under a new model with the consequence that a much, much larger joint probability table would have to be constructed. Just one assigned probability Q_i for,

$$P(X_1, X_2, \cdots, X_{N+M} \,|\, \mathcal{M}_k)$$

could now reflect a dependency on what had happened on all of the preceding trials. But I don't think we care to contemplate the immediate combinatorial explosion that such a "solution" offers.

I prefer not to go down this path for a different reason. I think it better to try and identify a small number of *causal factors* $U, V, W, \cdots$ behind the truth of X at any trial where the rationale for the numerical assignment is once again the information inserted via model $\mathcal{M}_k$. Then, some joint probability, say,

$$P(X_t = x_1, U_t = u_2, V_t = v_1, W_t = w_2 \,|\, \mathcal{M}_k)$$

can appear on the right hand side under the summation. I believe this approach is better than postulating probabilities conditioned on previous appearances of X via some sort of Markov chain claiming that these previous appearances in and of themselves exert some causal influence at further trials.

Of course, we still have to deal with a larger joint probability table when we start to include causal factors, but the trade–off is that we get to retain independence at different trials as well as the comfort of a well–defined and unchanging state space.

12.3 Bruno de Finetti's Representation Theorem

At this juncture, I just want to present a very preliminary and cursory overview of de Finetti's Theorem since it has arisen in our previous discussion of this Chapter's topic on extending the formal manipulation rules. Basically, this section attempts nothing more than showing the correspondence between my preferred notation and other versions you will come across in the literature.

What it comes down to is that de Finetti's Theorem, no matter how it might appear in the various idiosyncratic notations employed, is essentially the same as what I have labeled as the prior predictive formula. In my notation, this is written as $P(M_1, M_2, \cdots, M_n)$.

More specifically, I will present versions of de Finetti's Theorem as given by Feller, Bernardo & Smith, Geisser, and, finally, a version given by Jaynes, all in their preferred notation. The intent here is to discern whether we can do any pattern matching among all of these varied expressions and thereby reduce some of the mystery surrounding this theorem.

Supplemental Exercise 12.3.1: Engage in a recapitulation of my preferred notation for the probability of a statement's first occurrence.

Solution to Supplemental Exercise 12.3.1:

This exercise is nothing but a repetition of the arguments presented in several places throughout my books. I like to start out with the most primitive of formal manipulation rules which justifies why an IP assigns a value of 1/2 to the probability for the first appearance of HEADS in the coin flipping scenario, or a value of 1/6 to the first appearance of, say, the THREE face in the dice rolling scenario.

The numerical assignment to the probability of observing the ith observation for statement A is first an application of the **Sum Rule**,

$$P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i, \mathcal{M}_k)$$

and secondly an application of the **Product Rule**,

$$P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

The first term indicates that the actual numerical value assigned as a probability is dependent on the assumption of the truth of some model $\mathcal{M}_k$. The information contained in model $\mathcal{M}_k$ and inserted into the probability distribution for A thus determines whether $P(A = a_i \,|\, \mathcal{M}_k)$ is assigned a value of 0, 1/2, 1, or anything in between.

The second term is the prior probability for the models and indicates the IP's initial degree of belief in any particular assignment in the first term. It would be completely acceptable within this framework to substitute arbitrary, but still legitimate, numbers, say, $P(A = a_i \,|\, \mathcal{M}_{59}) = 0.27$, and $P(\mathcal{M}_{59}) = 0.84$.

However, if Laplace's supposition of an IP who is "totally uninformed about the physical causes" for an observation is adopted as a standard baseline, the prior probability over model space is "flat." The consequence is that the first appearance

of an observation has probability $1/(M+n-1) = 1/n$. For example, the probability of observing the THREE face up on the first roll of an unknown die is,

$$P(A = \text{THREE}) = \frac{1}{M+n-1} = \frac{1}{1+6-1} = \frac{1}{6}$$

where $M = 1$ represents the first appearance of an observation, and $n = 6$ is the dimension of the arbitrarily defined state space for the roll of a die.

Supplemental Exercise 12.3.2: Engage in a recapitulation of my preferred notation for the prior predictive probability of future frequency counts.

Solution to Supplemental Exercise 12.3.2:

The argument in the previous exercise can be extended to any number M of future observations. We will now revert back to the simple coin tossing scenario of $n = 2$ because the introductory version of de Finetti's Theorem is restricted to that case.

Place a subscript on statement A to indicate the future trial. The prior predictive probability for the first three observations, $M = 3$, then becomes,

$$P(A_3, A_2, A_1) = \sum_{k=1}^{\mathcal{M}} P(A_3 \,|\, A_2, A_1, \mathcal{M}_k) \times P(A_2 \,|\, A_1, \mathcal{M}_k) \times P(A_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

Since, as part of the problem definition, $n = 2$, the probability for each term depends only on the information in model $\mathcal{M}_k$ and not on the results of any previous observations. Thus the above can be simplified to,

$$P(A_3, A_2, A_1) = \sum_{k=1}^{\mathcal{M}} P(A_3 \,|\, \mathcal{M}_k) \times P(A_2 \,|\, \mathcal{M}_k) \times P(A_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

Suppose the IP is interested in the sequence of HEADS, TAILS, HEADS on the first three coin flips. Adopt a short–hand notation of,

$$q = P(A = \text{HEADS} \,|\, \mathcal{M}_k) \text{ and } (1-q) = P(A = \text{TAILS} \,|\, \mathcal{M}_k)$$

then,

$$P(A_3 = a_1, A_2 = a_2, A_1 = a_1) = \sum_{k=1}^{\mathcal{M}} q \times (1-q) \times q \times P(\mathcal{M}_k)$$

Advance to the first generalization for any M, where $M = M_1 + M_2$. M_1 is the number of observations a_1, while M_2 is the number of observations a_2,

$$P(A_M = a_1, A_{M-1} = a_2, \cdots, A_1 = a_1) = \sum_{k=1}^{\mathcal{M}} q^{M_1} \times (1-q)^{M_2} \times P(\mathcal{M}_k)$$

and then to the second generalization for a continuous prior probability over model space,

$$P(A_M = a_1, A_{M-1} = a_2, \cdots, A_1 = a_1) = \int_0^1 q^{M_1} \times (1-q)^{M_2} \times \text{pdf}\,(q)\ \text{d}q$$

and, finally, to the third generalization utilizing the multiplicity factor and pertaining to some macro–statement asking about the probability of M_1 and M_2 future frequency counts,

$$P(M_1, M_2) = \int_0^1 W(M)\, q^{M_1}\, (1-q)^{M_2}\ \text{pdf}\,(q)\ \text{d}q \tag{12.2}$$

The above is my version of de Finetti's Theorem.

Supplemental Exercise 12.3.3: As a rather important aside, does the framework permit any deviation from the adoption of the independence of past data assumption in the previous exercise?

Solution to Supplemental Exercise 12.3.3:

Yes it does. Suppose that the state space had not been defined with dimension $n = 2$ with the only possible observations on each trial as HEADS or TAILS. Suppose instead that the state space had been defined as $n = 8$ with possibilities HEADS, HEADS, HEADS, $\cdots$ TAILS, TAILS, TAILS over an aggregation of three consecutive tosses. Then, any model $\mathcal{M}_k$ would have specified information assigning numerical values to probabilities in an eight cell joint probability table. The IP would then have just one q_i value in a changed expression,

$$P(A_3 = a_1, A_2 = a_2, A_1 = a_1) = \int \cdots \int_{\sum_{i=1} q_i = 1} q_i\ \text{pdf}\,(q_1, q_2, \cdots, q_8)\ \text{d}q_i$$

Supplemental Exercise 12.3.4: Write out Feller's version of de Finetti's Theorem.

Solution to Supplemental Exercise 12.3.4:

On page 228 of Volume II [6], Feller has this expression for de Finetti's Theorem,

$$P\{S_n = k\} = \binom{n}{k} \int_0^1 \theta^k\, (1-\theta)^{n-k}\ F\{d\theta\}$$

Pattern matching, we conclude that $n \equiv M$, $k \equiv M_1$, $n - k \equiv M - M_1 = M_2$. The binomial expression $\binom{n}{k}$ is the multiplicity factor $W(M)$. The probability expression on the left hand side $P\{S_n = k\}$ is then translated into the probability for M_1 observations from a total of M frequency counts. Implicitly, Feller is also finding the probability for $M_2 \equiv n - k$ frequency counts as well.

The dimension of the state space is implicitly understood in Feller's presentation to consist of only two possible observations. Feller prefers binary "random variables" expressed as $X = 1$ or $X = 0$ instead of my *statements* $(A = a_1)$ or $(A = a_2)$.

Feller's "parameters" θ and $(1 - \theta)$ are seen to be the same as the assigned numerical values q and $(1-q)$ under some model. The rather strange looking symbol $F\{d\theta\}$ is Feller's preferred very abstract measure theory notation for a cumulative probability distribution. But, in the end, he also uses a *beta distribution* probability density function when it comes time for an example.

Earlier, in his Equation (6.1) on page 55, Feller had offered up a more transparent version of de Finetti's theorem looking like this,

$$P\{S_n = k\} = \binom{n}{k} \int_0^1 p^k \, (1-p)^{n-k} \, u(p) \, dp \qquad k = 0, \cdots, n$$

However, the unfortunate choice of the symbol "p" to indicate a "parameter" on the right hand side can be confused with the symbol "P" on the left hand side to indicate a "probability." Feller confuses us even more when he also labels p as a "random variable." Feller uses the language of "randomizing a binomial distribution" as the intuitive meaning for this expression.

In his *Example (a)*, he even works out an explicit example for Laplace's "totally uninformed IP" (my characterization, of course, not Feller's) by integrating by parts with $u(p) = 1$. His solution in the form of the easily recognizable $(n+1)^{-1}$ is confirmation that he was using a *beta distribution* with parameters $\alpha = \beta = 1$ for $u(p)$, the "mixing" distribution.

Notice my correct usage of the word "parameters" to refer to the parameters of the *beta distribution*, whereas Feller uses a parameter p to refer to q, an assigned numerical value as conditioned on the truth of the information in some model.

Nonetheless, Feller would agree with Bayes, Laplace, Jaynes, and me that the probability for seeing five HEADS in nine future flips of the coin is 1/10 when "randomizing the binomial distribution" according to a particular prior probability for the "parameter p." For example, in the case of $n = 9$ flips of the coin, and $k = 5$ "successes," that is, with "success" defined as an appearance of HEADS,

$$P\{S_9 = 5\} = (n+1)^{-1} = \frac{1}{10}$$

My way of computing the same answer as Feller derives from the prior predictive formula of,

$$P(M_1 = 5, M_2 = 4) = \int_0^1 W(M) \, q^{M_1} \, (1-q)^{M_2} \, C_{\text{Beta}} \, q^{\alpha-1} \, (1-q)^{\beta-1} \, \mathrm{d}q$$

Immediately after the above solution to this example of an application of de Finetti's Theorem, Feller tells us that,

> In gambling language (6.1) corresponds to the situation where a skew coin is picked by a chance mechanism and then trials are performed with this coin of unknown structure. To a gambler the trials do not look independent; indeed if a long sequence of heads is observed it becomes likely that for our coin p is close to 1 and so it is safe to bet on further occurrences of heads.

Think about this for a bit. The intuitive analogy Feller is proposing should really be to *many* biased coins that will be picked with no favoritism from some huge "bag" of biased coins. The bias in the coins runs the gamut from two tailed coins, coins heavily biased to TAILS, $\cdots$, fair coins, $\cdots$, coins heavily biased to HEADS, all the way through to two headed coins.

The procedure follows this protocol. Suppose that the first pick from the bag is a coin biased to TAILS with $P(A = \text{HEADS} \,|\, \mathcal{M}_k) = 0.25$. With a *beta distribution* representing total ignorance about the coins, this assignment of $Q = 0.25$ came about because any q had an equal chance of being selected. In other words, this coin biased towards TAILS was an outcome from the huge bag of biased coins reflecting the reality of Feller's $u(p) = 1$ or my $P(\mathcal{M}_k) \equiv \text{pdf}\,(q) = 1$. The degree of belief that five HEADS and four TAILS *would be seen* in nine future coin flips with this coin is then,

$$P\{S_9 = 5\} = P(M_1 = 5, M_2 = 4) = \binom{9}{5} (1/4)^5 \, (3/4)^4 = 0.0389$$

But here is the critical conceptual point that Feller missed. He says that "and then trials are performed with this coin of unknown structure ..." Conceptually, no trials are performed. If they had been, then the problem would have transitioned into a posterior predictive probability, in other words, a calculation of probability using the formula for $P(M_1, M_2 \,|\, N_1, N_2)$ instead of $P(M_1, M_2)$. The probability calculation is for the degree of belief, *if* nine trials of the coin flip *had* taken place with this biased coin, that the outcome *would have been* five HEADS and four TAILS.

There can be no long sequence of HEADS actually observed, there can only be a probability for any sequence of HEADS as calculated from the particular model's assignment of a numerical value to the probability for HEADS. And these assignments must, by definition, cover the entire spectrum from 0 to 1 with an equal relative weighting. This is what, of course, $u(p) = 1$ dictates.

The predictive probability expression tells us that there must be an integration over q. This corresponds to picking another biased coin from the bag. Suppose this newly chosen coin is heavily biased to HEADS. This corresponds, for example, to the "correct" assignment according to $\mathcal{M}_k \to (q, 1-q) \to (0.95, 0.05)$. Remember there is an equal chance of picking out any q along the line from 0 to 1 from a prior probability $P(\mathcal{M}_k) \to$ "flat". Now, the degree of belief that five HEADS and four TAILS *would be seen* in nine future coin flips with this coin is then,

$$P\{S_9 = 5\} = P(M_1 = 5, M_2 = 4) = \binom{9}{5} (0.95)^5 \, (0.05)^4 = 0.0006$$

This all becomes very clear from an operational standpoint because it is the same procedure as when we try to approximate an exact integration through a Monte Carlo approach. So, as a back of the envelope verification of the result of a probability of 1/10 for seeing five HEADS and four TAILS as provided by either Feller's example of de Finetti's Theorem, or my prior predictive formula, construct this *Mathematica* function,

```
f[n_, k_, start_, end_, inc_] :=
   NumberForm[Total[Table[Binomial[n, k] q^k (1 - q)^(n - k),
   {q, start, end, inc}]] / ((end - start) / inc + 1), {5, 4}]
```

An average of 19 binomial probabilities covering the range from $q = 0.05$ to $q = 0.95$ in increments of 0.05,

```
f[9, 5, .05, .95, .05]
```

computes to $P(M_1 = 5, M_2 = 4) \approx 0.1053$, while a finer gradation over a longer range of q, for an average of 999 binomial probabilities,

```
f[9, 5, .001, .999, .001]
```

computes to $P(M_1 = 5, M_2 = 4) \approx 0.1001$.

Supplemental Exercise 12.3.5: Write out Bernardo & Smith's version of de Finetti's Theorem.

Solution to Supplemental Exercise 12.3.5:

On page 174, as **Corollary 1**, Bernardo & Smith give us their version of de Finetti's Theorem in the form of,

$$p(x_1 + \cdots + x_n = y_n) = \int_0^1 \binom{n}{y_n} \theta^{y_n} (1-\theta)^{n-y_n} \, dQ(\theta)$$

This is similar to Feller's version with the x_i called "0–1 random quantities" instead of statements. The y_n are Feller's k and my M_1. Their n is the same as Feller's n and the same as my $M = M_1 + M_2$. They also persist in calling θ a "parameter."

Their verbal rationale for de Finetti's Theorem is not too different than most other interpretations including mine, except for one glaring discordancy.

> The interpretation of this representation theorem is of profound significance from the point of view of subjectivist modelling philosophy. It is *as if*:
> (i) the x_i are judged to be independent $\cdots$ conditional on a random quantity θ;
> (ii) θ is itself assigned a probability distribution Q;
> (iii) by the strong law of large numbers, $\theta = \lim_{n\to\infty}(y_n/n)$, so that Q may be interpreted as "beliefs about the limiting relative frequency of 1's".

There is absolutely no reason that makes sense to me as to why θ has to be interpreted as the limit of the number of successes divided by the number of trials as the number of trials goes to infinity. θ comes from a prior probability distribution as they readily admit to in their expression. So, as I tried to demonstrate in the previous exercise, the operation is first, a selection of a θ from the prior probability, second, a computation of the binomial probability with this θ, and ended with an integration over all values of θ from 0 to 1.

Thus, I see no qualification placed on θ because θ is supposed to assume one limiting value as the number of trials go to infinity. The IP doesn't have to imagine any kind of limiting procedure; it's completely superfluous. All the θs it will ever require are right there in the prior probability.

Supplemental Exercise 12.3.6: Write out Geisser's version of de Finetti's Theorem.

Solution to Supplemental Exercise 12.3.6:

Geisser's presentation [9, pp. 72–74, Example 3.5] was very helpful to me in sorting out my own derivations of both the prior and posterior predictive formulas. I deconstructed the example accompanying his version of the theorem very closely and learned a lot in the process of doing so. Despite the expected differences in personal notation, the pattern follows all of the other versions of de Finetti's Theorem.

> *Theorem.* To every infinite sequence of exchangeable random binary variables $\{W_k\}$ corresponds a probability distribution $F_U(u)$ concentrated on [0, 1] such that for every permutation of the integers $1 \cdots M$ and any given integer $r \in [0, M] \ldots$
> b. $\Pr[\sum_{k=1}^{M} W_k = r] = \binom{M}{r} \int_0^1 u^r \, (1-u)^{M-r} \, dF(u)$
> c. $\lim_{M\to\infty} M^{-1} \sum_{k=1}^{M} W_k = U$

Almost every word at the beginning of this theorem has no meaning for me. I want to calculate the probability for a finite sequence of occurrences, not an infinite sequence. I don't use the language of binary random variables; I set up a state space of dimension $n = 2$ where there is only one statement A. It can be observed as true for only two possibilities, $P(A = a_1 \mid \mathcal{M}_k)$ or $P(A = a_2 \mid \mathcal{M}_k)$. My degree of belief in the truth of the statement $(A = a_i)$ is dependent on the assumption of the truth of a model $\mathcal{M}_k$ whose information has provided a numerical value $Q_i \equiv P(A = a_i \mid \mathcal{M}_k)$ for my degree of belief.

Furthermore, I do not require any reliance on the concept of some permutation of exchangeable random variables. There is instead the expression at any trial t,

$$P(A_t = a_1 \mid A_{t-1}, \cdots, A_1, \mathcal{M}_k) = Q_1$$

because it is only the model over the defined state space that provides the numerical value of the probability, and not the presence of previous observations. Likewise,

at some different time t',

$$P(A_{t'} = a_2 \,|\, A_{t'-1}, \cdots, A_1, \mathcal{M}_k) = Q_2$$

for the same reason. If there are M_1 occurrences of $(A = a_1)$ and M_2 occurrences of $(A = a_2)$, this leads to the appearance of the terms $q^{M_1}\,(1-q)^{M_2}$.

By the way, this is a curious way of saying that the formal manipulation rules (and de Finetti's Theorem) demolish the core notion behind *time series*. The time t at which something occurred is not a causal factor. The particular trial at which a success or failure will take place doesn't make any difference to an IP's degree of belief in the total number of successes and failures.

Despite all of this quibbling, Geisser's final formula corresponds exactly to my prior predictive formula. However, just like Bernardo & Smith, and everyone else, Geisser felt compelled to add the unnecessary condition that the average of the number of successes in the limit of an infinite number of trials must equal some true value of Q_1.

The mention of de Finetti's Theorem is followed by an informative example. However, Geisser's Example 3.5 actually works out to an expression for the posterior predictive probability, rather than giving an example, as he should have, for the prior predictive probability. So, my solution in these supplemental exercises more naturally belongs to the content domain for Chapter Fourteen. I will, in fact, treat it there in full detail, but for present purposes relating to de Finetti's Theorem, let's examine the first part of Example 3.5.

Here is the first part of Geisser's Example 3.5 on page 73,

> Let X_i be independent random variables and $\Pr(X_i = 1) = \theta$, then $T = \sum X_i$ with probability
>
> $$\binom{N}{t}\,\theta^t\,(1-\theta)^{N-t}$$
>
> Suppose
>
> $$p(\theta) = \frac{\Gamma(\alpha+\beta)\,\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\,\Gamma(\beta)}$$

Matching this up with my framework and notation, the problem is translated into this familiar language. The state space has dimension $n = 2$ consisting of the two statements, $(A = a_1)$ and $(A = a_2)$, or even more explicitly, 1) "The coin being flipped will show HEADS.", and 2) "The coin being flipped will show TAILS."

Assign a numerical value to the probability that statement $(A = a_1)$ is TRUE conditioned on assuming the truth of a model $\mathcal{M}_k$ through the familiar expression $P(A = a_1 \,|\, \mathcal{M}_k) = q$. I use q here instead of Q_1 because I want to emphasize the eventual integration over model space. The complementary assigned numerical value to the probability that statement $(A = a_2)$ is TRUE conditioned on assuming the truth of a model $\mathcal{M}_k$ is expressed as $P(A = a_2 \,|\, \mathcal{M}_k) = (1-q)$.

The IP is interested in computing its degree of belief for any number of future frequency counts for the statements $(A = a_1)$ and $(A = a_2)$. There will be M_1 future counts for statement $(A = a_1)$ and M_2 future counts for statement $(A = a_2)$. Together, $M_1 + M_2 = M$, the total number of times the coin will be flipped.

Thus, the IP is interested in computing, say, the probability of $M_1 = 5$ future HEADS and $M_2 = 4$ future TAILS from a total of $M = 9$ future coin flips. This is $P(M_1, M_2)$, the prior predictive probability. Take note that, as far as the IP is concerned, this coin has never been flipped before so there is no past data on the occurrences of HEADS or TAILS.

Because the assigned numerical value of the probability depends only on the model defined over the state space, and not on any previous outcomes, that is,

$$P(A_t = a_1 \mid A_{t-1}, \cdots, A_1, \mathcal{M}_k) = P(A_t = a_1 \mid \mathcal{M}_k) = q$$

any invocation of the **Product Rule** over some number of trials always results in a term,

$$q^{M_1} \times (1-q)^{M_2}$$

Furthermore, there are $\binom{M}{M_1}$ different ways that M_1 HEADS and M_2 TAILS could occur over the M trials with each one of these ways characterized by the above term. For example, if interest centers on the probability of obtaining two HEADS and one TAIL in three future flips of the coin, then $M = 3$, $M_1 = 2$, $M_2 = 1$, $\binom{3}{2} = 3$, and the resultant first term in the prior predictive formula becomes,

$$3 \times q^2 \times (1-q)^1 \rightarrow \binom{M}{M_1} q^{M_1} (1-q)^{M_2}$$

Explicitly, the first way is $q \times q \times (1-q)$, the second way is $q \times (1-q) \times q$, and the third way is $(1-q) \times q \times q$. The first way is $P(A_3 = a_1) \times P(A_2 = a_1) \times P(A_1 = a_2)$, the second way is $P(A_3 = a_1) \times P(A_2 = a_2) \times P(A_1 = a_1)$, and the third way is $P(A_3 = a_2) \times P(A_2 = a_1) \times P(A_1 = a_1)$. Every one of these expressions should have been conditioned on the model as in $P(A_t = a_i \mid \mathcal{M}_k)$, but I hope you appreciate the desire to unclutter.

Geisser's expression in the above quote,

$$\binom{N}{t} \theta^t (1-\theta)^{N-t}$$

is the analog to my development. His N is my M, the total number of future frequency counts; his t is my M_1, the total number of "successes"; his $N - t$ is my M_2, the total number of "failures"; his θ and $(1-\theta)$ are my q and $(1-q)$, a model's purported "correct" assignment of a legitimate numerical value for the probabilities of a success or a failure.

Even though we might have deleted statements $(A_{t-1}, \cdots, A_1)$ on the right hand side of the conditioned upon symbol (the "solidus"), we can never get rid of the

statement $\mathcal{M}_k$. From the simple and generic formal manipulation rules,

$$P(A) = \sum_{k=1}^{\mathcal{M}} P(A, \mathcal{M}_k) = \sum_{k=1}^{\mathcal{M}} P(A \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

the right hand side of the prior predictive formula must look like,

$$P(M_1, M_2) = \sum_{k=1}^{\mathcal{M}} \binom{M}{M_1} q^{M_1} \, (1-q)^{M_2} \, P(\mathcal{M}_k)$$

$P(\mathcal{M}_k)$ is the notorious "prior probability" that seems to be the source of so much angst among Bayesians. For the life of me, I don't understand what all the fuss is about. It is a very well behaved, non–mysterious, proper probability that captures the information impacting the degree of belief in all of the $\mathcal{M}$ models. Recall that these statements $\mathcal{M}_k$ took the form of: "The correct assignment of a numerical value of probability to the statement $(A = a_1)$ is q." An IP's degree of belief that this higher–level statement, shown wrapped in quotes, is TRUE is simply $P(\mathcal{M}_k)$.

Since q can take on a continuous range of values between 0 and 1 inclusive, a probability density function is required. For the beginning case of $n = 2$, a *beta distribution* is a practical implementation for $P(\mathcal{M}_k)$. Because the *beta distribution* is *conjugate* to the first binomial term, an analytical solution for the predictive probability formulas becomes tractable. Like all MEP distributions, the information in the *beta distribution* is reflected in its parameters, here in the traditional notation of α and β.

After substituting this probability density function into the equation for the prior predictive formula, we now have,

$$P(M_1, M_2) = \binom{M}{M_1} \int_0^1 q^{M_1} \, (1-q)^{M_2} \, \mathrm{pdf}\,(q) \, \mathrm{d}q \tag{12.3}$$

Thus, my analog to Geisser's expression $dF(u)$ in de Finetti's Theorem is almost exactly the same as his when he writes out his $p(\theta) \equiv \mathrm{pdf}\,(q)$ as,

$$p(\theta) = \frac{\Gamma(\alpha+\beta)\, \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\Gamma(\alpha)\, \Gamma(\beta)}$$

Now, the prior predictive formula in Equation (12.3) looks like,

$$P(M_1, M_2) = \binom{M}{M_1} \int_0^1 q^{M_1} \, (1-q)^{M_2} \, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\, \Gamma(\beta)} \, q^{\alpha-1} (1-q)^{\beta-1} \, \mathrm{d}q \tag{12.4}$$

In this form, it is ready for the usual series of transformations,

$$
\begin{aligned}
P(M_1, M_2) &= \binom{M}{M_1} \int_0^1 q^{M_1}\,(1-q)^{M_2}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,q^{\alpha-1}(1-q)^{\beta-1}\,\mathrm{d}q \\
&= \binom{M}{M_1}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^1 q^{M_1}\,(1-q)^{M_2}\,q^{\alpha-1}(1-q)^{\beta-1}\,\mathrm{d}q \\
&= \binom{M}{M_1}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^1 q^{M_1+\alpha-1}(1-q)^{M_2+\beta-1}\,\mathrm{d}q \\
&= \binom{M}{M_1} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \times \frac{\Gamma(M_1+\alpha)\,\Gamma(M_2+\beta)}{\Gamma(M+\alpha+\beta)} \qquad (12.5)
\end{aligned}
$$

We could stop right here and implement Equation (12.5) in *Mathematica* with an alternative version of the prior predictive formula,

```
ppProb[M_, M1_, alpha_, beta_] := Module[{t1, t2, t3},
        t1 = Binomial[M, M1];
        t2 = Gamma[alpha + beta] / (Gamma[alpha] Gamma[beta]);
        t3 = (Gamma[M1 + alpha] Gamma[M - M1 + beta]) /
                                    Gamma[M + alpha + beta];
        N[t1 t2 t3]]
```

Apply this to our current running example of the prior predictive probability of observing five HEADS in nine coin flips,

```
ppProb[9, 5, 1, 1]
```

which returns the correct probability of 0.10.

This answer is predicated on setting $\alpha = \beta = 1$ for the final two arguments of **ppprob[ ]**. That is, the IP's degree of belief in the correct q is distributed equally from 0 to 1. Colloquially, it is fair to describe this situation as an IP who is in a state of total ignorance about the physical characteristics of the coin that will be flipped, as well as all of the details as to how it will be flipped. Geisser also substitutes $\alpha = \beta = 1$ for his solution at the end of his example.

Unfortunately, as I mentioned at the outset, Geisser's purported example of de Finetti's Theorem is confusing on first exposure because, instead of deriving as I just did a *prior* predictive formula, he sets out to derive a *posterior* predictive formula. To accomplish that objective, he needs $P(\mathcal{M}_k \,|\, \mathcal{D})$, or in his notation, $p(\theta \,|\, x^{(N)})$. Starting out from our preferred template and then switching over to Geisser's notation, we have first,

$$
P(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)} \rightarrow \frac{\textbf{Numerator}}{\textbf{Denominator}}
$$

From this point on, the derivation follows Geisser,

$$p(\theta \,|\, x^{(N)}) \equiv P(\mathcal{M}_k \,|\, \mathcal{D})$$

$$\textbf{Numerator} \;= \binom{N}{t} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, \theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}$$

$$\textbf{Denominator} \;= \int_0^1 \binom{N}{t} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, \theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}\; d\theta$$

$$p(\theta \,|\, x^{(N)}) = \frac{\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}}{\int_0^1 \theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}\; d\theta}$$

$$\int_0^1 \theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}\; d\theta = \frac{\Gamma(t+\alpha)\,\Gamma(N-t+\beta)}{\Gamma(N+\alpha+\beta)}$$

$$p(\theta \,|\, x^{(N)}) = \frac{\Gamma(N+\alpha+\beta)\;\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}}{\Gamma(t+\alpha)\,\Gamma(N-t+\beta)}$$

Supplemental Exercise 12.3.7: Write out Jaynes's version of de Finetti's Theorem.

Solution to Supplemental Exercise 12.3.7:

I have saved Jaynes's version for last. Somewhat surprisingly, he takes a different tack than evidenced in the more orthodox presentations of de Finetti's Theorem as exemplified in the previous exercises. On page 586 of his book, Equation (18.86) is Jaynes's version of de Finetti's Theorem,

$$P(n \,|\, N) = \int_0^1 \mathrm{d}p\, p^n (1-p)^{N-n} g(p)$$

First of all, Jaynes is the first to have a conditional probability of the left hand side. While one couldn't say that this way of writing out the expression is wrong, I still don't like it. Despite the fact that an IP is allowed to put anything to the right of the conditioned upon symbol to explicitly indicate to the reader whatever is assumed as given, I don't like it mainly because I think the appearance of the conditioned upon symbol should be used to highlight the importance of whether data is or is not present in the probability expression. For de Finetti's Theorem and the prior predictive probability, the data are not present. Thus, it might be confusing to see an expression conditioned on a symbol N.

Also, I do not like that Jaynes uses the symbol "p" on the right hand side and the symbol "P" on the left hand side just as Feller did in his alternate version of de Finetti's Theorem. It is confusing that $p \equiv P(A_t = a_i \,|\, \mathcal{M}_k)$, or God forbid, $\theta \equiv P(A_t = a_i \,|\, \mathcal{M}_k)$, whereas $P(n \,|\, N)$ is something else again.

Jaynes's n is my M_1 and his N is my $M = M_1 + M_2 \equiv N = n + (N - n)$. No one, in any version of de Finetti's Theorem, disputes the essential criterion of an integration over q (Jaynes's p) from 0 to 1 in the binary case. For Jaynes, the relative weighting in that integration over p is reflected by $g(p)$.

Far more interesting is that Jaynes calls our prior probability over model space $P(\mathcal{M}_k)$, a "generating function" $g(p)$, a well known concept going back to Laplace in more rigorous mathematical treatments of probability theory. Immediately after the presentation of Equation (18.86), Jaynes delves into a very complicated discussion of how he envisions the proof of de Finetti's Theorem proceeding.

Jaynes's proof is very condensed and touches on many advanced topics in quick succession, but I think there's something very important there. I intend to take it up in a different place because it demands a huge amount of expansion.

Or should I say, in order for my brain to understand it, there must be an unwinding of the macro. It also appears to me that, at least on the surface, there is some similarity between Jaynes's technical proof and Feller's. For example, both mention that the problem reduces to the *Hausdorff moment problem*.

Something else that Jaynes says that is easy to understand in comparison with the rest of his development is that the generating function $g(p)$ is exactly the same as his "A_p distribution." This is a critical confirmation for me since I have already explained at some length (Section 32.8 of Volume II) why Jaynes's A_p distribution is the same as my prior probability over model space $P(\mathcal{M}_k)$. They would have to be in order for any mapping to take place between my prior predictive formula and Jaynes's version of de Finetti's theorem. It's satisfying for Jaynes to come right out and flatly assert their equivalency.

Chapter 13

Predicting College Success

13.1 Feller's Sample Space

The beginning exercises supplement the extensive discussion of elementary points, the sample space, and events as introduced in Chapter Thirteen. There I chose to adopt Feller's abstract language that talked about placing r balls into n cells.

My Chapter Thirteen discussed Feller's pedagogical example of events and a sample space as presented in his Chapter 1, Volume 1, pp. 7–14 [6] where $n = 3$ and $r = 3$. The beginning exercise examines the easy extension to $n = 3$ and $r = 4$. In other words, we are examining the placement of four balls into three cells, or less abstractly, $M = 4$ repetitions from a state space of dimension $n = 3$.

The second of these beginning exercises re–examines one that Feller himself presented [6, pg. 125, Vol 1] concerning the throw of two dice. In this case, $n = 6$ and $M = 2$, or the placement of two balls into six cells. For some reason, Feller did not choose to solve this easy problem using his own sample space approach, but I will reproduce his answer using this tactic.

I am particularly proud of my effort here in section 13.3 in deconstructing Feller's canonical *Probability for the number of accidents in a week* that appeared in his discussion of occupancy problems. I'm sure there must be many people like me who have labored over the numbers appearing in Feller's Table 1 on his page 40.

Their decryption surprisingly follows the detailed explanation first proffered in Chapter Thirteen, Volume I. The first two sections in these supplemental exercises recapitulate the general line of thinking required before delving into clearing up Feller's Table 1.

I then finish up with exercises on Feller's **Birthday Problem**, Bose–Einstein and Fermi–Dirac statistics from statistical mechanics, and conclude with pre–data inferences about college students.

Supplemental Exercise 13.1.1: Discuss the sample space arising from the placement of four balls into three cells.

Solution to Supplemental Exercise 13.1.1:

The sample space represents an abstract notion. The $n = 3$ cells refer to any three statements in a state space. Thus, it might refer to a coin toss where the possible measurements are HEADS, TAILS, or EDGE. It might refer to a dice roll where the possible outcomes of ONE and TWO are labeled as FIRST, the possible outcomes of THREE and FOUR are labeled as SECOND, and the possible outcomes of FIVE and SIX are labeled as THIRD. It might refer to the three results of a test labeled as HIGH, AVERAGE, and LOW.

The $r = 4$ balls to be placed into the $n = 3$ cells are labeled as **a, b, c, d**. This labeling means that the balls are supposed to be "distinguishable."

I prefer to use the notation of M in place of r to refer to future frequency counts. $M = 4$ might refer to the four separate tosses of the coin at four distinct times where **a** is the first toss, **b** is the second toss, **c** is the third toss, and **d** is the fourth toss. It might refer to four differently colored dice in the dice roll where **a** is the red die, **b** is the blue die, **c** is the green die, and **d** is the yellow die. It might refer to four different students who took some test where **a** is Alex, **b** is Beth, **c** is Carl, and **d** is Dawn. It might refer to M different "harmonic oscillators" in statistical mechanics where the n cells are the permissible energy states for each oscillator.

There are $n^r \equiv n^M = 3^4 = 81$ *elementary points* in the sample space for this situation.[1] Thus, there can be fifteen different possible contingency tables,

$$\frac{(M+n-1)!}{M!\,(n-1)!} = \frac{(4+3-1)!}{4!\,2!} = 15$$

Three examples of the fifteen possible contingency tables are $\boxed{4}\boxed{0}\boxed{0}$, $\boxed{0}\boxed{2}\boxed{2}$, and $\boxed{1}\boxed{2}\boxed{1}$ illustrating some of the frequency counts that might ensue when $M = 4$ balls have been distributed over $n = 3$ cells.

The first contingency table might chronicle the frequency counts in a coin toss where HEADS appeared on all four tosses. The second contingency table might hold the frequency counts for two dice that were observed as SECOND and two that were observed as THIRD. The third contingency table might reflect the frequency counts where one student scored HIGH, two students scored AVERAGE, and one student scored LOW on a test.

This total of fifteen contingency tables can be broken down further. There are three tables where $M = 4$ is the sum $4 + 0 + 0$. There are six tables where $M = 4$ is the sum $3 + 1 + 0$. There are three tables where $M = 4$ is the sum $2 + 2 + 0$. And finally there are also three tables where $M = 4$ is the sum $2 + 1 + 1$. So we see that indeed, $3 + 6 + 3 + 3 = 15$.

[1] Feller makes an error here. On page 10, he says that there are $4^3 = 64$ elementary points.

The number of tables mentioned above can be found from the formula,

$$\textbf{Number of tables} = \frac{n!}{r_z!\, r_s!\, r_d!\, r_t! \cdots r_M!}$$

For example, for the sum $3 + 1 + 0$, there is one repetition of the zero count, one repetition of a single count, no repetition of a double count, one repetition of a triple count, and no repetitions of a quadruple count, leading to,

$$\frac{n!}{r_z!\, r_s!\, r_d!\, r_t!\, r_M!} = \frac{3!}{1!\, 1!\, 0!\, 1!\, 0!} = 6$$

If we take into account the "distinguishability" feature, then each one of these four sums is multiplied by its multiplicity factor $W(M)$. When we add up this multiplication over the four patterns, we find we reproduce the total number of 81 elementary points in the sample space.

It is obvious that the $4 + 0 + 0$ pattern can happen in only one way even after taking notice of the labeling **a, b, c, d** for the balls, coins, dice, or students. Thus, the first term counting towards the total number of elementary points is 3×1.

Each $3+1+0$ pattern can happen in four ways. For example, if the contingency table were $\boxed{0}\boxed{1}\boxed{3}$, ball **a** could be in the second cell, or ball **b** could be in the second cell, or ball **c** could be in the second cell, or ball **d** could be in the second cell, with the remaining three balls all in cell 3. We now have a partial sum consisting of $(3 \times 1) + (6 \times 4)$ contributing to the total number of elementary points.

The $2 + 2 + 0$ pattern can happen in six ways. We now have a partial sum consisting of $(3\times1)+(6\times4)+(3\times6)$ contributing to the total number of elementary points.

The final pattern of $2 + 1 + 1$ can happen in twelve different ways. We find this number, as well as all of the above numbers, through the multiplicity factor,

$$W(M) = \frac{M!}{M_1!\, M_2!\, M_3!} = \frac{4!}{2!\, 1!\, 1!} = 12$$

We now have all four terms,

$$(3 \times 1) + (6 \times 4) + (3 \times 6) + (3 \times 12) = 3 + 24 + 18 + 36 = 81 = n^M = 3^4$$

adding up to the correct total number of elementary points.

Any elementary point can now be easily interpreted within the context of these different hierarchies. For example, consider the situation where four students have taken a test where the score on the test can only be HIGH, AVERAGE, or LOW. This scenario is an example of a sample space with a total of 81 elementary points.

$M = 4$ is the number of students, or abstractly, $r = 4$ balls. $n = 3$ represents the number of statements in the state space, which consists of "The student scored HIGH on the test.", "The student scored AVERAGE on the test.", and "The student

scored LOW on the test.", or, abstractly, $n = 3$ cells. Thus, there are, in total, $n^r = n^M = 3^4 = 81$, elementary points in this sample space.

At the highest level, one possible decomposition of the sum $M = 4$ happens to be $2 + 1 + 1$. A possible contingency table adhering to this breakdown is $\boxed{1}\boxed{2}\boxed{1}$. One student scored HIGH, two students scored AVERAGE, and one student scored LOW according to this selected contingency table out of the three possible. There are two other possible contingency tables, and they are $\boxed{2}\boxed{1}\boxed{1}$ and $\boxed{1}\boxed{1}\boxed{2}$.

Table 13.1 below summarizes the above remarks. It also shows that the total number of elementary points can be found as a sum over the multiplicity factors at each possible integer partitioning of the sum $M = 4$,

$$\sum_{j=1}^{\frac{(M+n-1)!}{M!\,(n-1)!}} W_j(M) \longrightarrow \sum_{j=1}^{15} W_j(M) = (3 \times 1) + (6 \times 4) + (3 \times 6) + (3 \times 12) = n^M = 81$$

Table 13.1: *The decomposition of the total sum of 81 possible elementary points for the case where $M = 4$ balls are distributed over $n = 3$ cells.*

Listing j	Contingency Tables	Number of Possibilities	Multiplicity Factor	Elementary Points
1	$\boxed{4}\boxed{0}\boxed{0}$	3	1	3
2	$\boxed{0}\boxed{4}\boxed{0}$			
3	$\boxed{0}\boxed{0}\boxed{4}$			
4	$\boxed{3}\boxed{1}\boxed{0}$	6	4	24
5	$\boxed{3}\boxed{0}\boxed{1}$			
6	$\boxed{1}\boxed{3}\boxed{0}$			
7	$\boxed{1}\boxed{0}\boxed{3}$			
8	$\boxed{0}\boxed{1}\boxed{3}$			
9	$\boxed{0}\boxed{3}\boxed{1}$			
10	$\boxed{2}\boxed{2}\boxed{0}$	3	6	18
11	$\boxed{2}\boxed{0}\boxed{2}$			
12	$\boxed{0}\boxed{2}\boxed{2}$			
13	$\boxed{2}\boxed{1}\boxed{1}$	3	12	36
14	$\boxed{1}\boxed{2}\boxed{1}$			
15	$\boxed{1}\boxed{1}\boxed{2}$			
Total		15		81

There are twelve possible ways that the four students could have produced the contingency table $\boxed{1}\boxed{2}\boxed{1}$. This is also true for the other two contingency tables. Proceeding down to the lowest level of each elementary point, one of these twelve ways is that Beth scored HIGH, Alex and Dawn scored AVERAGE, and Carl scored LOW. In terms of Feller's r balls and n cells, balls **a** and **d** were distributed to cell 2, ball **b** was distributed to cell 1, and ball **c** was distributed to cell 3.

Supplemental Exercise 13.1.2: Solve Feller's example concerning the throw of two dice with a straightforward sample space solution.

Solution to Supplemental Exercise 13.1.2:

Here is how Feller presented the problem: [6, pg. 125, Vol 1]

> Two true dice are thrown. The events "ace with first die" and "even face with second" are independent since the probability of their simultaneous realization, $\frac{3}{36} = \frac{1}{12}$, is the product of their probabilities, namely $\frac{1}{6}$ and $\frac{1}{2}$.

The sample space consists of $n^M = 6^2 = 36$ elementary points. The depiction of an elementary point in the sample space shows six cells with each cell representing a possible realization of one of the six statements from the state space. In other words, did a die show the ONE face (the ace), the TWO face (an EVEN face), and so on? Two balls labeled as **a** and **b** represent a red die and a green die to capture "distinguishability."

After all thirty six points in this sample space are listed below in Figure 13.1, we can pick out the particular elementary points that satisfy the events Feller describes. For example, at elementary point #19, the red die shows the ONE face and the green die the FOUR face.

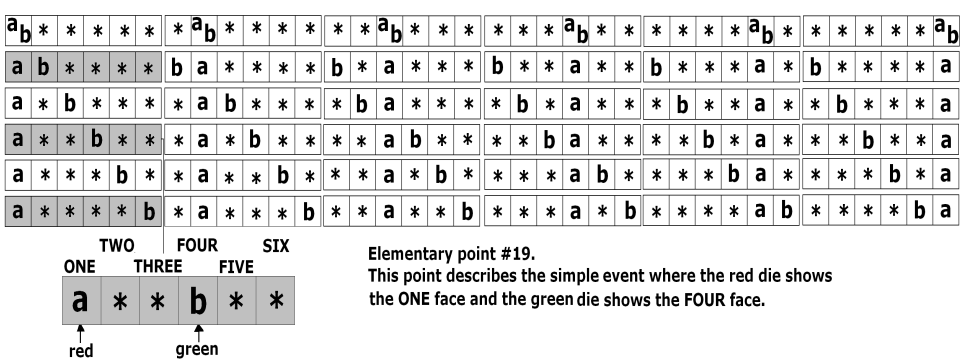

Figure 13.1: *All 36 elementary points in the sample space for Feller's dice example.*

The first row shows the six contingency tables where the same face appears on both dice. The remaining thirty contingency tables in the last five rows show where

the dice have different faces. The elementary points satisfying the criteria for the compound event must come from these thirty contingency tables.

Let's declare that the elementary points are labelled by reading across each of the six rows. Scanning the entire list of elementary points, only points 7, 19, and 31 satisfy the criteria of a ONE face on the red die and a TWO or FOUR or SIX face on the green die. These three elementary points in the sample space are marked in gray. Since there are only three of these points, by Feller's (imperfect) definition of a probability where each elementary point has the same probability (the die is "true"), the probability of this event is,

$$P(\textbf{Event}) = \frac{\text{points satisfying event}}{n^M} = \frac{3}{36} = \frac{1}{12}$$

13.2 Simple and Compound Events

Feller's definition of an *event*, divided into simple and compound events, can be explained through an inferential scenario where we roll the dice. For this particular exercise, let $M = 5$ and $n = 6$. In other words, one die is rolled five times in succession. The dimension of the state space is $n = 6$, where the statements are the familiar $(A = a_1) \rightarrow$ "The die showed the ONE spot." through $(A = a_6) \rightarrow$ "The die showed the SIX spot.".

In Feller's terminology, $r = 5$ balls, **a** through **e**, have been distributed over $n = 6$ cells. That five balls are "distinguishable" means either that five differently colored dice were rolled once, or, as in our case here, one die was rolled at five separately identifiable times. The total number of elementary points in the sample space is $n^r \equiv n^M = 6^5 = 7,776$.

The total number of 252 possible contingency tables is found through,

$$\frac{(M+n-1)!}{M!\,(n-1)!} = \frac{(5+6-1)!}{5!\,5!} = 252$$

An example of one of these 252 contingency tables is $\boxed{0}\boxed{2}\boxed{0}\boxed{2}\boxed{0}\boxed{1}$ representing two TWOs, two FOURs, and one SIX in the five rolls of the die. The multiplicity factor $W(M)$ for this contingency table is,

$$W(M) = \frac{5!}{0!\,2!\,0!\,2!\,0!\,1!} = 30$$

telling us that there are thirty different ways this contingency table could have come about when taking into account the particular time order in which the die faces appeared. The only SIX might have been occurred on the first roll, the first FOUR at the second roll, the first TWO at the third roll, the second TWO at the fourth roll, and the second FOUR on the fifth and final roll.

Now examine the partitioning of the integer sum $M = 5$. The listing of all seven partitionings is,

Partition 1: $5 + 0 + 0 + 0 + 0 + 0 = 5$

Partition 2: $4 + 1 + 0 + 0 + 0 + 0 = 5$

Partition 3: $3 + 2 + 0 + 0 + 0 + 0 = 5$

Partition 4: $3 + 1 + 1 + 0 + 0 + 0 = 5$

Partition 5: $2 + 2 + 1 + 0 + 0 + 0 = 5$

Partition 6: $2 + 1 + 1 + 1 + 0 + 0 = 5$

Partition 7: $1 + 1 + 1 + 1 + 1 + 0 = 5$

The number of tables at each partitioning is found from the formula,

$$\textbf{Number of tables} = \frac{n!}{r_z!\, r_s!\;\, r_d!\; \cdots\; r_M!}$$

For example, for the partitioning $5 + 0 + 0 + 0 + 0 + 0$, there are five repetitions of the zero count, no repetition of a single count through a quadruple count, and one repetition of the five count, leading to,

$$\frac{n!}{r_z!\, r_s!\, r_d!\; \cdots\; r_M!} = \frac{6!}{5!\; 0!\; 0!\; 0!\; 0!\; 1!} = 6$$

This result that there must be a total of six contingency tables with all five frequency counts in some cell of the contingency table obviously is the right answer.

How many tables are there where there are two frequency counts for one face, two frequency counts for another face, and a count of one for the final face? This situation is represented by the fifth integer partition, $2 + 2 + 1 + 0 + 0 + 0 = 5$, as they were enumerated above. This actual example was presented above in the contingency table $\boxed{0}\boxed{2}\boxed{0}\boxed{2}\boxed{0}\boxed{1}$. Application of the above formula finds that there are 60 such contingency tables,

$$\frac{n!}{r_z!\, r_s!\, r_d!\; \cdots\; r_M!} = \frac{6!}{3!\; 1!\; 2!\; 0!\; 0!\; 0!} = 60$$

In how many different ways can this particular contingency table happen given the specification that the "balls are distinguishable," or, in other words, given that the die was rolled successively at five separate and distinguishable times? Applying the formula for the multiplicity factor,

$$W(M) = \frac{M!}{M_1!\; M_2!\; \cdots\; M_n!} = \frac{5!}{0!\; 2!\; 0!\; 2!\; 0!\; 1!} = 30$$

we find that there are 30 different ways for this particular configuration of frequency counts to occur. This all leads to a contribution of $60 \times 30 = 1800$ elementary points to the overall total of 7,776 elementary points in the sample space.

All of these computations are summarized below in Table 13.2. Importantly, the total of 252 different possible contingency tables and the total of 7,776 elementary points in the sample space are confirmed.

Table 13.2: *The decomposition of the total sum of 7776 possible elementary points for the case where one die is rolled five times.*

Partition Number	Integer Partition	Number of Possibilities	Multiplicity Factor	Elementary Points
1	$5+0+0+0+0+0$	6	1	6
2	$4+1+0+0+0+0$	30	5	150
3	$3+2+0+0+0+0$	30	10	300
4	$3+1+1+0+0+0$	60	20	1200
5	$2+2+1+0+0+0$	60	30	1800
6	$2+1+1+1+0+0$	60	60	3600
7	$1+1+1+1+1+0$	6	120	720
Total		252		7776

With all of this tedious groundwork accomplished, it is easy to give examples of Feller's *events* for this sample space. A *simple event* is any one of the 7,776 elementary points in the sample space. From the discussion above, we know that there are 60 possible contingency tables documenting the fact that the same face appeared twice, and three other different faces each appeared just once. This is the occupancy pattern represented by the integer partition $2+1+1+1+0+0=5$, listed as the sixth partition in Table 13.2.

Let's select one of these 60 possibilities as the contingency table $\boxed{0}\boxed{1}\boxed{2}\boxed{1}\boxed{1}\boxed{0}$ representing two THREEs, and one TWO, FOUR, and FIVE face appearing in the five rolls of the die. Now, this particular contingency table can also happen in 60 different ways. Select one of these 60 possibilities as $\boxed{\star}\boxed{\mathbf{c}}\boxed{\mathbf{b,\ d}}\boxed{\mathbf{e}}\boxed{\mathbf{a}}\boxed{\star}$ which represents the two THREEs as occurring on the second and fourth rolls, the TWO on the third roll, the FOUR on the fifth and last roll, and the FIVE on the first roll.

This is one example of how a simple event **A** is defined. It can happen in only one way. Move up one level to a compound event **B**. Compound event **B** might be defined as "the same face appears on all five rolls." This must be represented by the first integer partition, which can happen in six ways: ONE appeared on all five rolls, $\cdots$, SIX appeared on all five rolls. A more involved compound event **C** might

be defined as "any situation except that happenstance where each face appears just once." This event includes the first six integer partitions, but excludes the seventh partition. This compound event **C** would then be composed of 7056 elementary points from the total of 7776 elementary points in the entire sample space.

Events might also be defined by the sum over the five faces that appeared during the rolls. For an easy event to figure out, suppose event **D** is when the sum of the faces is 28 or greater. We first surmise that the possible range of the sum over five rolls must lie between 5 and 30 inclusive.

The sum of 30 is that elementary point in the sample space represented by $\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{5}$, in other words, a SIX on all five rolls. This is an example of one of the six possibilities of the first partition. A sum of 29 would be $\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{1}\boxed{4}$, in other words, a SIX on four rolls and a FIVE on some single roll. This is an example of one of the thirty possibilities of the second partition. Whereas the sum of 30 could happen in only one way, a sum of 29 can happen in five ways as we observe in the fourth column of Table 13.2. We are up to six elementary points that define the event **D**.

Finally, the sum of 28 would be $\boxed{0}\boxed{0}\boxed{0}\boxed{1}\boxed{0}\boxed{4}$, in other words, a SIX on four rolls and a FOUR on some single roll. This is another example of one of the thirty possibilities of the second partition. A sum of 28 can also happen in five ways from the multiplicity factor. We have reached the final accounting where there are eleven elementary points that define the event **D**.

The critical conceptual error that Feller made was in how he defined a probability assignment to events. He took each elementary point in the sample space to have a probability of n^{-r}. For our current example, each elementary point would have a probability assignment of $6^{-5} = 1/7776$. The probability of any event is then the number of elementary points defining the event divided by n^r, the total number of elementary points in the sample space. The language that Feller would employ sounds like this: *The* probability *of* event **D** *is* 11/7776.

But we now know that such an assignment of n^{-r} is derived from just *one* conceivable assignment of $1/n$ to the probability that any face would appear on any given roll. The MEP will make the numerical assignment of $1/n = 1/6$ at each trial under the information from the fair model, $P(A_t = a_i \,|\, \mathcal{M}_{\text{Fair}}) = 1/6$. The formal rules governing probability manipulations would then dictate,

$$P(A_5 \,|\, \mathcal{M}_{\text{Fair}}) \times P(A_4 \,|\, \mathcal{M}_{\text{Fair}}) \times \cdots \times P(A_1 \,|\, \mathcal{M}_{\text{Fair}}) = (1/6)^5 \equiv n^{-r}$$

Following Feller's mode of reasoning, accumulate this probability over all of the elementary points defining an event to find the probability for the event. For example, *the* probability *of* event **C** *is* $P(\mathbf{C}) = 1 - (720/7776)$.

13.3 Probability of Accidents During the Week

Here are four supplementary exercises that present a detailed solution to the classical occupancy problem concerning the number of accidents distributed across the days of the week.

Feller solves this problem on pages 39–40 and Table 1 as his,

Example. (c) *Configurations of $r = 7$ balls in $n = 7$ cells.*

I mentioned this problem in a footnote during Exercise 13.8.7 of Volume I.

These supplemental exercises are merely a continuation of the discussion of elementary points and the sample space as broached in the previous sections, and as it was presented originally in my section 13.7 of Volume I. The culmination of the exercises also serve as confirmation of how Feller was going to define *the* probability *of* an event as the aggregation of the number of elementary points defining the event over n^r.

Supplemental Exercise 13.3.1: Discuss in a laconic manner the main ingredients in this occupancy problem.

Solution to Supplemental Exercise 13.3.1:

How are a number of accidents distributed across the days of the week? Feller solves for the specific case of seven accidents. It might have made things clearer for the reader if Feller had chosen a number different than 7 to disambiguate the number of accidents from the seven days of the week. Nonetheless, we solve here the $r = 7$ and $n = 7$ case to see if we can duplicate Feller's results.

So, for Feller, $r = 7$ balls represent seven accidents, and $n = 7$ cells represents the seven days of the week. In our notation, $r = 7$ becomes $M = 7$ for the number of frequency counts of future accidents. Feller uses n to refer abstractly to "cells," where we use n to refer to the number of statements in the state space.

As always, $n = 7$ is the dimension of the state space. If we employ our standard notation, the state space consists of seven statements, $(A = a_1)$ through $(A = a_7)$. $(A = a_1)$ is the statement, "An accident occurred on Monday.", $(A = a_2)$ is the statement, "An accident occurred on Tuesday.", $\ldots$, $(A = a_7)$ is the statement, "An accident occurred on Sunday.".

The $r = 7$ "distinguishable balls" are the $M = 7$ "distinguishable" accidents. Thus, we might label them as simply, the first accident, the second accident, and so forth in order to make them distinguishable. This is just like a temporal ordering for coin flips, color of the die in dice tossing, or names of students or kangaroos. This level of detail is imposed in order to implement distinguishability. When we say that

the "balls are indistinguishable," we mean that we voluntarily discard information about which accident occurred first, second, or last.

The sample space consists of $n^M = 7^7 = 823,543$ elementary points. The total number of possible contingency tables containing the frequency counts is,

$$\textbf{Contingency tables} = \frac{(M+n-1)!}{M!\,(n-1)!} = \frac{13!}{7!\,6!} = 1,716$$

The fifteen integer partitions of the sum $M = 7$ are:

Partition # 1: 7+0+0+0+0+0+0

Partition # 2: 6+1+0+0+0+0+0

Partition # 3: 5+2+0+0+0+0+0

Partition # 4: 5+1+1+0+0+0+0

Partition # 5: 4+3+0+0+0+0+0

Partition # 6: 4+2+1+0+0+0+0

Partition # 7: 4+1+1+1+0+0+0

Partition # 8: 3+3+1+0+0+0+0

Partition # 9: 3+2+2+0+0+0+0

Partition # 10: 3+2+1+1+0+0+0

Partition # 11: 3+1+1+1+1+0+0

Partition # 12: 2+2+2+1+0+0+0

Partition # 13: 2+2+1+1+1+0+0

Partition # 14: 2+1+1+1+1+1+0

Partition # 15: 1+1+1+1+1+1+1

Feller calls these integer partitions the *Occupancy numbers* in the first column of his Table 1. He presents them in reverse order of the above.

An example of the first occupancy number indicating that all seven accidents happen on the same day of the week is that all the accidents happen on Wednesday. An example of the second occupancy number that six accidents happen on the same day of the week and one accident happens on a different day is that six accidents happen on Tuesday and one accident happens on Friday.

An elementary point is the most detailed statement we can make. This is where we take distinguishability into account. Consider the third occupancy number where

five accidents happen on the same day of the week and two accidents happen on a different day. An elementary point belonging to this event is that the second, third, fourth, sixth and seventh accidents happen on Thursday and the first and fifth accidents happen on a Sunday.

One possible contingency table from the 1,716 possible frequency counts is,

0	2	3	0	0	1	1

illustrating an example of the tenth partitioning of $M = 7$ as they have just been listed. This contingency table is a future frequency count for seven accidents where no accidents happen on Monday, two accidents happen on Tuesday, three accidents happen on Wednesday, no accidents happen on Thursday or Friday, and one accident happens on both Saturday and Sunday.

Supplemental Exercise 13.3.2: Use the counting formula to decompose the total number of contingency tables.

Solution to Supplemental Exercise 13.3.2:

In the first exercise above, we found that there were a total of 1,716 different possible frequency counts for seven accidents to occur over a week. The first term in the counting formula from Chapter Thirteen allowed us to decompose this total into the sum of its constituent parts. In other words, for each of the fifteen occupancy numbers, find the number of different ways the pattern can be satisfied. The formula is,

$$\textbf{Contingency tables per occupancy number} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!}$$

For the previous example of the tenth listed occupancy number 3+2+1+1+0+0+0 there are,

$$\frac{7!}{3!\ 2!\ 1!\ 1!\ 0!\ 0!\ 0!\ 0!} = 420$$

different examples following this pattern, one of which is $\boxed{0}\boxed{2}\boxed{3}\boxed{0}\boxed{0}\boxed{1}\boxed{1}$.

An easy result from this formula that can be comprehended immediately is,

$$\frac{7!}{6!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 1!} = 7$$

telling us that there are seven different examples of the first listed integer partition 7+0+0+0+0+0+0. One possibility is the frequency count $\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{7}$ where all accidents happen on Sunday. Another easy result is that the least numerous occupancy number consists of just the one contingency table where one accident happens on every day of the week. These are the first and last entries in the table below.

Table 13.3 on the next page presents this calculation carried out for each one of the fifteen occupancy patterns, or class of contingency tables.

Table 13.3: *The decomposition of the total sum of 1,716 possible contingency tables.*

Occupancy Pattern	Occupancy Number	Number of Possibilities
1	$7+0+0+0+0+0+0$	7
2	$6+1+0+0+0+0+0$	42
3	$5+2+0+0+0+0+0$	42
4	$5+1+1+0+0+0+0$	105
5	$4+3+0+0+0+0+0$	42
6	$4+2+1+0+0+0+0$	210
7	$4+1+1+1+0+0+0$	140
8	$3+3+1+0+0+0+0$	105
9	$3+2+2+0+0+0+0$	105
10	$3+2+1+1+0+0+0$	420
11	$3+1+1+1+1+0+0$	105
12	$2+2+2+1+0+0+0$	140
13	$2+2+1+1+1+0+0$	210
14	$2+1+1+1+1+1+0$	42
15	$1+1+1+1+1+1+1$	1
	Total	1716

Supplemental Exercise 13.3.3: Complete the decomposition of the total number of elementary points in the sample space.

Solution to Supplemental Exercise 13.3.3:

The above counting exercise is only half the battle. Every one of these contingency tables can occur in the number of ways indicated by the multiplicity factor when taking into account that the accidents are distinguishable. In this manner, we can count up all of the elementary points that compose the sample space.

For example, the most numerous occupancy number is the tenth occupancy pattern which can come about in 420 different ways. However, each one of these contingency tables can itself occur in 420 different ways when taking the multiplicity factor and distinguishability into account,

$$W(M) = \frac{M!}{M_1! \times M_2! \times \cdots \times M_7!} = \frac{7!}{3!\ 2!\ 1!\ 1!\ 0!\ 0!\ 0!\ 0!} = 420$$

This illustrates why it would have helped to use a number of accidents different than the number of the days of the week. The calculation of the second term in the counting formula comes up with the same number as the first term. So it is confusing. Nonetheless, there are $420 \times 420 = 176,400$ elementary points in the sample space that define this event.

Seven different contingency tables cover the 7+0+0+0+0+0+0 integer partition. This describes the situation where all seven accidents occur on the same day of the week. But there is only one way for all seven accidents to occur on any one day, even when accounting for the fact that the accidents are distinguishable given that we know the time of their occurrence,

$$W(M) = \frac{M!}{M_1! \times M_2! \times \cdots \times M_7!} = \frac{7!}{7!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!} = 1$$

Only $7 \times 1 = 7$ elementary points are contributed to the overall sum for this case.

There are seven ways that each one of the, as Feller calls them, 6+1+0+0+0+0+0 occupancy numbers can occur,

$$W(M) = \frac{M!}{M_1! \times M_2! \times \cdots \times M_7!} = \frac{7!}{6!\ 1!\ 0!\ 0!\ 0!\ 0!\ 0!\ 0!} = 7$$

so $42 \times 7 = 294$ elementary points are contributed to the overall sum for this case.

Suppose that $\boxed{0}\boxed{0}\boxed{0}\boxed{1}\boxed{6}\boxed{0}\boxed{0}$ is one of the 42 possibilities for this occupancy pattern. Six accidents happen on Friday and one accident on Thursday. But the very first accident could occur on Thursday and the remaining six accidents on Friday, or the second accident could occur on Thursday and the first and third through seventh accidents on Friday, or, ..., or the last accident could occur on Thursday and the first six accidents could occur on Friday.

Table 13.4 on the next page shows the calculation of all 823,543 elementary points in the sample space. The first column lists the fifteen occupancy numbers. The second column is the first term of the counting formula which calculates the different number of contingency tables for each occupancy number.

The third column is the second term in the counting formula which calculates the multiplicity factor, that is, the number of different ways that each one of the contingency tables can occur given that the accidents are distinguishable. The final column is the multiplication of these two terms in the counting formula to provide the partial sum contributed by a particular occupancy pattern. The grand total at the bottom of the table shows the total number of contingency tables and the total number of elementary points in the sample space.

Here is an example of one of these 823,453 elementary points. Start out at the highest level of the thirteenth occupancy pattern where the sum $M = 7$ is decomposed into 2+2+1+1+1+0+0. This describes no accidents on two days, one accident on three days, and two accidents on two days.

There are 210 possible frequency counts fitting this pattern. Suppose that the frequency count $\boxed{1}\boxed{2}\boxed{1}\boxed{2}\boxed{0}\boxed{0}\boxed{1}$ is one of these 210 possibilities. Two accidents happen on Tuesday and Thursday, one accident happens on Monday, Wednesday, and Sunday, and no accidents happen on Friday and Saturday.

Now move down to the lowest level where we take notice of the time order of the particular accidents. There are 1260 different ways that the above contingency

Table 13.4: *The decomposition of the total sum of 823,543 elementary points in the sample space.*

Occupancy Pattern	First term	Second term	Partial total
$7+0+0+0+0+0+0$	7	1	7
$6+1+0+0+0+0+0$	42	7	294
$5+2+0+0+0+0+0$	42	21	882
$5+1+1+0+0+0+0$	105	42	4410
$4+3+0+0+0+0+0$	42	35	1470
$4+2+1+0+0+0+0$	210	105	22050
$4+1+1+1+0+0+0$	140	210	29400
$3+3+1+0+0+0+0$	105	140	14700
$3+2+2+0+0+0+0$	105	210	22050
$3+2+1+1+0+0+0$	420	420	176400
$3+1+1+1+1+0+0$	105	840	88200
$2+2+2+1+0+0+0$	140	630	88200
$2+2+1+1+1+0+0$	210	1260	264600
$2+1+1+1+1+1+0$	42	2520	105840
$1+1+1+1+1+1+1$	1	5040	5040
Totals	1716		823543

table could occur if the IP doesn't voluntarily discard this information. Here is a statement that is an elementary point, "The first accident happens on Wednesday, the second accident on Sunday, the third accident on Monday, the fourth accident on Thursday, the fifth accident also on Thursday, and finally, the sixth and seven accidents both on Tuesday. "

Supplemental Exercise 13.3.4: Tie in the developments so far from these exercises with Feller's Table 1.

Solution to Supplemental Exercise 13.3.4:

The first column in Feller's Table 1, page 40, is the listing of the occupancy numbers in the reverse order in which we have presented them. The second column is an application of the counting formulas in a somewhat confusing manner.

In the column heading, Feller says that $7! \times 7!$ is divided by the entries in the column. These numbers come from the numerators in the first and second counting formulas, in other words, they are $n!$ and $M!$.

The entry in the column associated with each occupancy pattern is then the elements in the denominator of the two formulas. For example, look at the first row

for the occupancy pattern 1+1+1+1+1+1+1. The counting formulas reveal that,

$$\frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} \times \frac{M!}{M_1! \times M_2! \times \cdots \times M_n!}$$

The numerator in the two terms is always going to be $7! \times 7!$ for all fifteen occupancy patterns. The denominator is,

$$(0! \times 7! \times 0! \times 0! \times 0! \times 0! \times 0! \times 0!) \times (1! \times 1! \times 1! \times 1! \times 1! \times 1! \times 1!) = 7! \times 1!$$

as Feller shows under the second column.

Now look at the more complicated entry under the second column in row 3. Here we have the occupancy pattern 2+2+1+1+1+0+0. The terms in the denominator are going to reduce down to,

$$(2! \times 3! \times 2! \times 0! \times 0! \times 0! \times 0! \times 0!) \times (2! \times 2! \times 1! \times 1! \times 1! \times 0! \times 0!) = 2!\ 3!\ 2! \times 2!\ 2!$$

Thus, even though Feller cryptically says in the column heading that this is the number of arrangements, he is actually calculating in the second column the partial contribution to the total number of elementary points in the sample space. I hope you will agree with me that this objective was accomplished more transparently in Table 13.4 of these exercises.

The most misleading aspect appears in the third column. Here, Feller presents *the* probability *of* each occupancy number as the total number of arrangements (that is, the number of elementary points in that event) divided by 7^7, or $n^r \equiv n^M$. For example, Feller lists a probability of 0.214197 in the third column for what is our tenth occupancy pattern, 3+2+1+1+0+0+0. Sure enough, the total number of elementary points for this event is 176,400 which when divided by n^M is equal to 0.214197.

But this is the probability only under one very specific model, namely the fair model where each elementary point has the same probability of n^{-M}. For all other assigned probabilities to the statements in the state space, this probability is wrong. For example, when the IP is totally uninformed about the causes of accidents, the probability for this occupancy pattern is,

$$\frac{420}{1716} = 0.244755$$

instead of 0.214197.

The probability for other events is much more discrepant. Our first occupancy pattern, where all the accidents occur on the same day of the week, has a probability of only 0.000008 according to Feller. But if the IP is "totally ignorant" about all

the models rather than completely certain about one model, then the probability for this event is instead,

$$\frac{7}{1716} = 0.004079$$

over 500 times larger.

It is interesting to comment on just the counting aspects of the problem where at least Feller's numbers here are correct even if we can't agree with him on the probability assignments.

The occupancy pattern, or event **E**, that can happen in the most number of ways is 2+2+1+1+1+0+0 comprising 264,600 elementary points in the sample space. Feller, of course, asserts that *the true probability* of this event is,

$$P(\mathbf{E}) = \frac{264,600}{823,543} = 0.321295$$

This pattern describes all those occupancy numbers with a single accident on three days of the week, two accidents on two days of the week, and no accidents on two days of the week.

The probability for the event of the occupancy pattern 1+1+1+1+1+1+1 where there is an accident every day of the week comprises only 5,040 elementary points. We might have selected this outcome as our intuitive choice of what could occur the maximum number of ways because it is the one "most evenly distributed." Anyone who mistakenly claims that such evenly distributed outcomes like this are "maximum entropy" solutions is seriously confused.

13.4 The Birthday Paradox

Another classic problem in probability that Feller addresses with his abstraction of r balls and n cells to define events in a sample space is called *the Birthday problem*, or, sometimes to emphasize the counter–intuitive answer arrived at, *the Birthday paradox*.

The problem is easy enough to state. How many people would have to be gathered together in a room in order for the probability for at least one match of birthdays to reach a reasonably high value, say, 2/3? The answer is surprising because the number of people required seems way too few. Here, the answer we will use for the numerical example is that a gathering of thirty people will result in a probability of about 70% for a matching birthday.

Supplemental Exercise 13.4.1: Solve the Birthday problem according to Feller's criteria.

Solution to Supplemental Exercise 13.4.1:

The 365 days in the year play the role of the n cells. The M number of people gathered together play the role of the r balls distributed over the n cells. Through trial and error, we can find that value of M that satisfies our specification of a high probability of a matching birthday. A typical statement in the state space $(A = a_i)$ would be something like, "A person has a birthday on May 12th."

The number of elementary points in this sample space is truly enormous at a value of $n^r \equiv n^M = 365^{30} \approx 7.39 \times 10^{76}$. We could begin to break down the sum $M = 30$ into its integer partitions as we have done before. But if we recognize that the partition involving everyone having a birthday on a different date is the only one we must compute, then it is simply one minus this number for the number of ways that at least one match occurs.

This first counting formula for the relevant integer partition of $M = 30$ looks like,

$$\overbrace{1 + \cdots + 1}^{30} + \overbrace{0 + \cdots + 0}^{335} = 30$$

with the number of different possible frequency counts,

$$\textbf{Contingency tables} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} = \frac{365!}{335!\ 30!\ 0!\ \cdots\ 0!}$$

The second counting formula involving the multiplicity factor is,

$$W(M) = \frac{30!}{1!\ \cdots 1!} = 30!$$

The number of elementary points comprising this event is then,

$$\textbf{Elementary points} = \frac{365!}{335!\ 30!} \times 30! = \frac{365!}{335!}$$

Feller's definition of the probability of an event over the sample space says that the probability of this event of no matching birthdays is,

$$\frac{365!/335!}{n^r} = \frac{365!/335!}{365^{30}}$$

This is the same as the formula Feller provides as Equation (3.1) on page 31 for the probability of no repetition,

$$p = \frac{(n)_r}{n^r} = \frac{n(n-1)\cdots(n-r+1)}{n^r}$$

Now, we can go ahead and compute the probability of the compound event of at least one match as,

$$P(\textbf{Match}) = 1 - \frac{365 \times 364 \times \cdots \times 336}{365^{30}} = 0.706316$$

It would seem plausible that the major contribution to the overall probability just given for at least one match would be the probability of exactly one match where two people have their birthday on the same day of the year, while the remaining 28 people all have birthdays on different days. Starting off with the same line of attack works here as well.

The particular integer partition of $M = 30$ of concern would be,

$$2 + \overbrace{1 + \cdots + 1}^{28} + \overbrace{0 + \cdots + 0}^{336} = 30$$

with the number of different possible frequency counts changing to,

$$\textbf{Contingency tables} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} = \frac{365!}{336!\, 28!\, 1!\, \cdots\, 0!}$$

The second counting formula involving the multiplicity factor changes to,

$$W(M) = \frac{30!}{2!\, 1!\, \cdots 1!} = \frac{30!}{2}$$

The number of elementary points comprising this new event is then,

$$\textbf{Elementary points} = \frac{365!}{336!\, 28!} \times \frac{30!}{2} = \frac{365!}{336!} \times 15 \times 29$$

The above is the new numerator for the probability of this event. The denominator remains the same as the total number of elementary points in the sample space n^r. The probability of exactly one matching birthday is then,

$$(365!/336! \times 15 \times 29)/365^{30} = 0.380216$$

It is interesting that a relatively large value of 0.32 of probability still remains to be accounted for in other kinds of matches. There might be exactly two matching birthdays and one would expect the probability of this event to be lower. How much of the missing probability does it deliver?

The relevant integer partition of $M = 30$ now becomes,

$$\overbrace{2 + 2}^{2} + \overbrace{1 + \cdots + 1}^{26} + \overbrace{0 + \cdots + 0}^{337} = 30$$

The probability for this event of two people with the same birthday one day of the year and another two people with the same birthday on another day is then,

$$\left(\frac{365!}{337!\, 26!\, 2!} \times \frac{30!}{2!\, 2!} \right) / 365^{30} = 0.213237$$

The two events of two people sharing one birthday together with two people sharing a birthday on two different days with a combined probability nearly 60% account for the bulk of the overall probability of at least one match.

Three sets of two people sharing the same birthday accounts for nearly another 7%. All of the other possible events with some form of a match make up the remaining 3%. For example, the probability that three people share the same birthday is,

$$\left(\frac{365!}{337!\,27!} \times \frac{30!}{3!}\right) / 365^{30} = 0.0105$$

There are highly improbable events that meet the criterion of a match and technically contribute to the final probability of 0.706316, but do so at an utterly negligible level. For example, what is the probability that 28 people from the 30 assembled have the same birthday? The particular integer partition of $M = 30$ of concern would now become,

$$28 + 1 + 1 + \overbrace{0 + \cdots + 0}^{362} = 30$$

with the number of different possible frequency counts changing to,

$$\textbf{Contingency tables} = \frac{n!}{r_z! \times r_s! \times r_d! \times \cdots \times r_M!} = \frac{365!}{362!\,2!\,0!\,\cdots\,1!\,\cdots\,0!}$$

The second counting formula involving the multiplicity factor changes to,

$$W(M) = \frac{30!}{28!\,2!\,\cdots\,0!}$$

The number of elementary points comprising this new event is then,

$$\textbf{Elementary points} = \frac{365!}{362!\,2!} \times \frac{30!}{28!\,2!}$$

The above is the new numerator for the probability of this event. The denominator remains the same as the total number of elementary points in the sample space n^r. The probability of 28 people having the same birthday is then 1.42×10^{-67}.

Supplemental Exercise 13.4.2: To lend additional weight to the findings of the last exercise, verify Feller's example (page 33) where a gathering of 23 people results in a probability slightly over one half for at least one match of birthdays.

Solution to Supplemental Exercise 13.4.2:

First, we just show the relevant integer partition of the new sum $M = 23$ where nobody has a common birthday,

$$\overbrace{1 + \cdots + 1}^{23} + \overbrace{0 + \cdots + 0}^{342} = 23$$

The number of elementary points comprising this event is then,

$$\textbf{Elementary points} = \frac{365!}{342!\,23!} \times 23! = \frac{365!}{342!}$$

Again, from Feller's definition of the probability of an event on a sample space, the probability of this event of no matching birthdays is,

$$\frac{365!/342!}{n^r} = \frac{365!/342!}{365^{30}}$$

The complementary event is one minus this probability,

$$P(\textbf{Match}) = 1 - \frac{365 \times 364 \times \cdots \times 343}{365^{30}} = 0.507297$$

Supplemental Exercise 13.4.3: Construct a *Mathematica* function to take over the computations for at least one match in the birthday problem.

Solution to Supplemental Exercise 13.4.3:

An easy very straightforward implementation is,

```
birthdayprob[npeople_] := 1 - N[(365! /(365 - npeople)!) /
                                    (365^npeople)]
```

13.5 Bose–Einstein Statistics

I based my discussion of the so–called "Bose–Einstein" statistics in section 13.7 of Volume I directly on Feller's example of discrete sample spaces. In his explanation of the probabilistic rationale behind Bose–Einstein statistics, which, by the way, is far better than any you will find in Physics or Chemistry textbooks, he continued, to our benefit, to rely upon the simple abstraction of r balls distributed over n cells. The first mention of Bose–Einstein statistics on pp. 11–12 and Table 2 used the language of "placing three *indistinguishable* balls into three cells."

Later on, in his **5. Applications to Occupancy Problems**, Chapter II, he picks up the thread that leads on once again to Bose–Einstein statistics. You will find my fully detailed deconstruction of Feller's **Examples** (a) appearing at the top of his page 42 with $n = 5$ and $r = 3$ in my Exercises 27.7.18 and 27.7.19 in Volume II as part of my treatment of Erwin Schrödinger's approach to statistical mechanics.

Feller dips into statistical mechanics as the appropriate context in which to begin his explanation of Bose–Einstein statistics (pp. 40–41).

> Consider a mechanical system of r indistinguishable particles. [Given my adoption of Schrödinger's language for statistical mechanics, this would be re–written instead as: *Consider a physical* **assembly** *of N indistinguishable* **systems**.] In statistical mechanics it is usual to subdivide the phase space into a large number, n, of small regions or cells so that each particle is assigned to one cell. In this way the state of the entire system is described in terms of a random distribution of the r particles in n cells. Offhand it would seem

> that (at least with an appropriate definition of the n cells) all n^r arrangements should have equal probabilities. If this is true, the physicist speaks of *Maxwell–Boltzmann statistics* ...
>
> Remember that we are here concerned only with *indistinguishable* particles. We have r particles and n cells. *By Bose–Einstein statistics we mean that only distinguishable arrangements are considered and that each is assigned probability* $1/A_{r,n}$ *with* $A_{r,n}$ *defined in (5.2)* ... [Emphasis in the original.]

Feller's Equation (5.2) which he refers to in the above quote is,

$$A_{r,n} = \binom{n+r-1}{r} = \binom{n+r-1}{n-1}$$

This is the same as my formula for the total number of contingency tables with M frequency counts distributed over the n cells of the contingency table,

$$\textbf{Total number of contingency tables} = \frac{(M+n-1)!}{M!\,(n-1)!}$$

His probability of $1/A_{r,n}$ for a "distinguishable arrangement" then becomes for me the probability for any frequency count under the information in $P(\mathcal{M}_k)$ reflecting lack of knowledge about physical causes,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!\,(n-1)!}{(M+n-1)!}$$

Feller almost employs this language when he says (bottom of page 41),

> To sum up: *the probability that cells number* $1, 2, \ldots, n$ *contain* $r_1, r_2, \ldots, r_n$ *balls, respectively (where* $r_1 + r_2 + \cdots + r_n = r$*) is given by* ... $1/A_{r,n}$ *under Bose–Einstein statistics;* ...

Supplemental Exercise 13.5.1: Recapitulate Feller's beginning example of a sample space as he used it to explain Bose–Einstein statistics.

Solution to Supplemental Exercise 13.5.1:

Feller presented in his Table 1, page 9, a complete listing of all 27 elementary points for the sample space where $r = 3$ and $n = 3$. On page 11, *(c) The case of indistinguishable balls.*, he says to consider the case where the $r = 3$ balls are indistinguishable. The 27 elementary points comprising Table 1 then reduce to 10 instances as shown in Table 2 on the next page. Where did the number 10 come from, and what is the relationship between Table 1 and Table 2?

There are, in fact, just 10 contingency tables for the sample space with $r = 3$ and $n = 3$. We know this from,

$$\textbf{Contingency tables} = \frac{(M+n-1)!}{M!\,(n-1)!} = \frac{5!}{3!\,2!} = 10$$

These are the ten frequency counts listed in Table 2. For example, the seventh instance listed in Table 2 is the frequency count $\boxed{1}\boxed{0}\boxed{2}$, where one ball is in the first cell, balls are absent from the second cell, and two balls are in the third and last cell. This would include three elementary points in the original sample space, namely points #13, #14, and #15 in Table 1.

It is easy to reconstruct the origin of all these numbers by following the mode of analysis used for all of these exercises. Table 13.5 shows the integer partition of the sum $M = 3$ in the first column. The second column is the value from the first counting formula, the contribution to the total number of contingency tables,

$$\textbf{Partial sum of contingency tables} = \frac{n!}{r_z!\, r_s!\, r_d!\, r_t!}$$

and the third column is the value from the multiplicity factor for each contingency table,

$$\textbf{Multiplicity factor} = \frac{M!}{M_1!\, M_2!\, M_3!}$$

The final column is the multiplication of the second and third columns showing the partial sum of the number of the elementary points in the sample space. The sum over the second column totals the number of instances where the balls are, as Feller says, "indistinguishable." The multiplicity factor doesn't count. The sum over the final column tallies up the total number of elementary points n^r. Thus, we have parsed out the relationship in the numbers appearing in Table 1 and Table 2.

Table 13.5: *The relationship between the elementary points in the sample space for $r = 3$ and $n = 3$ and the definition of Bose–Einstein statistics.*

Integer Partition of $M = 3$	Partial sum of contingency tables	Multiplicity formula	Partial sum of elementary points
$3+0+0$	3	1	3
$2+1+0$	6	3	18
$1+1+1$	1	6	6
Totals	10		27

When Feller gets around to the arguments justifying the probability assignments illustrating the difference between the Maxwell–Boltzmann assignment as compared to the Bose–Einstein assignment, he starts shuffling the pea under the shells. Your very close attention is demanded at this juncture.

Under **Examples.** *(b) Indistinguishable balls: Bose–Einstein statistics.* on page 20, he says that one could assign a probability of 1/9 to point 4 and 2/9 to point 10 in Table 2 if one stuck to the original rationale of assigning a probability of 1/27

to each elementary point in the sample space. Since point 4 is an example of the contingency table $\boxed{2}\boxed{1}\boxed{0}$, it has a multiplicity factor of 3 leading to,

$$P(4) = \frac{W(M)}{n^r} = \frac{3}{27} = 1/9$$

Likewise, since point 10 is an example of the contingency table $\boxed{1}\boxed{1}\boxed{1}$, it has a multiplicity factor of 6 leading to,

$$P(10) = \frac{W(M)}{n^r} = \frac{6}{27} = 2/9$$

However, for Bose–Einstein statistics, the probability of 1/10 is assigned to each of the ten distinguishable arrangements which must, of course, be the answer from $1/A_{r,n}$. In my language, the probability of 1/10 is the probability for any possible contingency table under "all encompassing ignorance,"

$$P(M_1, M_2, M_3) = \frac{M!\,(n-1)!}{(M+n-1)!} = \frac{3!\,2!}{5!} = 1/10$$

Supplemental Exercise 13.5.2: What is the absolutely vital conceptual distinction between me and Feller in arriving at this answer for Bose–Einstein statistics?

Solution to Supplemental Exercise 13.5.2:

First, I arrive at the answer by adhering strictly to the formal rules of probability manipulation. In other words, I simply follow the rules for finding the marginal probability of the future frequency counts. I do this by summing over all of model space. The conditional probability for the counts when conditioned on the assignment made under the information in the kth model $\mathcal{M}_k$ is multiplied by the prior probability for each model, and then summed,

$$P(M_1, M_2, M_3) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, M_3 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

Secondly, some rationale must be provided for the prior probability assignment. I accept Laplace's rationale that, unless otherwise specified, the IP is, in fact, totally uninformed about the causes leading to the appearance of a ball in a cell. When a non–informative prior probability is substituted for $P(\mathcal{M}_k)$, the answer returned by the formal manipulation rules of probability is indeed our desired value of 1/10.

Feller does not invoke this kind of rationale from the formal manipulation rules as outlined above in trying to justify the changed assignment from the Maxwell–Boltzmann situation to the Bose–Einstein situation. Instead, he indulges in an "argument by intimidation" where the notion of "indistinguishability" is at the forefront to finding the answer of 1/10. If one can find the correct answer by going through the ordinary steps already laid down in the conceptual foundations of

probability, why is it necessary to dredge up new concepts like "indistinguishability" that never before had made any appearance as part of the axiomatic foundation?

If there are no mysteries in my straightforward application of the rules, and I assert that none do exist, why drag in the very contentious mystery of what does it mean for a die, or a coin, or a kangaroo, or a student, or a particle to be indistinguishable? Indistinguishable by whom? Indistinguishable to the IP or to Nature herself? Einstein contributed to this confusion by arguing in defense of Bose's paper that "indistinguishability" of, say, photons was a justification of the result from physical principles.

Supplemental Exercise 13.5.3: Show in detail how the correct answer for "Bose–Einstein statistics" is arrived at by strictly following the formal manipulation rules.

Solution to Supplemental Exercise 13.5.3:

Start with the general probability expression given in the previous exercise, and then substitute the appropriate expressions for the first and second terms on the right hand side,

$$P(M_1, M_2, M_3) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, M_3 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

$$= \int \cdot \int_{\sum q_i = 1} \overbrace{W(M)\, q_1^{M_1}\, q_2^{M_2}\, q_3^{M_3}}^{\text{first term}}\, \overbrace{C_D\, q_1^{\alpha_1 - 1}\, q_2^{\alpha_2 - 1}\, q_3^{\alpha_3 - 1}}^{\text{second term}}\, \mathrm{d}q_i$$

The first term is recognized as the *likelihood* of the frequency counts under a specific model, while the second is the probability density function for the prior probability over model space. The instantiation of Laplace's uniform prior over model space will take place when the parameters of the Dirichlet probability density function are set at $\alpha_1 = \alpha_2 = \alpha_3 = 1$. The sum over a discrete number of models $\mathcal{M}$ has been transformed into a multiple integral over all legitimate values of q_i that sum to 1.

The next step recognizes those two terms that do not vary with the q_i, and therefore can be placed outside of the integration,

$$P(M_1, M_2, M_3) = W(M) \times C_D \times \int \cdot \int_{\sum q_i = 1} q_1^{M_1}\, q_2^{M_2}\, q_3^{M_3}\, q_1^{\alpha_1 - 1}\, q_2^{\alpha_2 - 1}\, q_3^{\alpha_3 - 1}\, \mathrm{d}q_i$$

$$= \frac{M!}{M_1!\, M_2!\, M_3!} \times \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1) \times \Gamma(\alpha_2) \times \Gamma(\alpha_3)} \quad \times$$

$$\int \cdot \int_{\sum q_i = 1} q_1^{M_1}\, q_2^{M_2}\, q_3^{M_3}\, q_1^{\alpha_1 - 1}\, q_2^{\alpha_2 - 1}\, q_3^{\alpha_3 - 1}\, \mathrm{d}q_i$$

Treat the integration in the third term on the right hand side separately,

$$\int \cdot \int_{\sum q_i=1} q_1^{M_1}\, q_2^{M_2}\, q_3^{M_3}\, q_1^{\alpha_1-1}\, q_2^{\alpha_2-1}\, q_3^{\alpha_3-1}\, \mathrm{d}q_i = \int \cdot \int_{\sum q_i=1} q_1^{M_1+\alpha_1-1}\, q_1^{M_2+\alpha_2-1}\, q_3^{M_3+\alpha_3-1}\, \mathrm{d}q_i$$

$$= \frac{\Gamma(M_1+\alpha_1)\times\Gamma(M_2+\alpha_2)\times\Gamma(M_3+\alpha_3)}{\Gamma[\,(M_1+\alpha_1)+(M_2+\alpha_2)+(M_3+\alpha_3)\,]}$$

Substitute this result for the third term to arrive at,

$$P(M_1,M_2,M_3) = \frac{M!}{M_1!\,M_2!\,M_3!}\times\frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\times\Gamma(\alpha_2)\times\Gamma(\alpha_3)}\times\frac{\Gamma(M_1+\alpha_1)\times\Gamma(M_2+\alpha_2)\times\Gamma(M_3+\alpha_3)}{\Gamma[\,(M_1+\alpha_1)+(M_2+\alpha_2)+(M_3+\alpha_3)\,]}$$

$$= \frac{M!}{M_1!\,M_2!\,M_3!}\times\frac{\Gamma(\sum_{i=1}^3\alpha_i)}{\Gamma(\alpha_1)\times\Gamma(\alpha_2)\times\Gamma(\alpha_3)}\times\frac{\Gamma(M_1+\alpha_1)\times\Gamma(M_2+\alpha_2)\times\Gamma(M_3+\alpha_3)}{\Gamma[\,\sum_{i=1}^3 M_i+\alpha_i\,]}$$

Labeling $\mathcal{A}=\sum_{i=1}^3\alpha_i$ and where, as always, $M=\sum_{i=1}^3 M_i$,

$$P(M_1,M_2,M_3) = \frac{M!}{M_1!\,M_2!\,M_3!}\times\frac{\Gamma(\mathcal{A})}{\Gamma(\alpha_1)\times\Gamma(\alpha_2)\times\Gamma(\alpha_3)}\times\frac{\Gamma(M_1+\alpha_1)\times\Gamma(M_2+\alpha_2)\times\Gamma(M_3+\alpha_3)}{\Gamma[\,M+\mathcal{A}\,]}$$

For any specific problem, M and $\mathcal{A}$ are fixed, so rearrange the above terms so that these constant values can appear in a separate constant term at the beginning,

$$P(M_1,M_2,M_3) = \frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})}\times\frac{\Gamma(M_1+\alpha_1)}{M_1!\,\Gamma(\alpha_1)}\times\frac{\Gamma(M_2+\alpha_2)}{M_2!\,\Gamma(\alpha_2)}\times\frac{\Gamma(M_3+\alpha_3)}{M_3!\,\Gamma(\alpha_3)}$$

Now, for the *pièce de résistance*! Please instantiate Laplace's uninformative prior probability over model space by setting the α_i parameters of the Dirichlet distribution to $\alpha_1=\alpha_2=\alpha_3=1$. Recalling that $M=3$ and $\mathcal{A}=\sum_{i=1}^3\alpha_i=3$, we substitute these values to find the probability for any pattern of future frequency counts to be our expected 1/10,

$$P(M_1,M_2,M_3) = \frac{3!\,\Gamma(3)}{\Gamma(6)}\times\frac{\Gamma(M_1+1)}{M_1!\,\Gamma(1)}\times\frac{\Gamma(M_2+1)}{M_2!\,\Gamma(1)}\times\frac{\Gamma(M_3+1)}{M_3!\,\Gamma(1)}$$

Since $\Gamma(M_i+1)=M_i!$ and $\Gamma(1)=1$, all three non–constant terms reduce to 1, leaving only the constant term,

$$P(M_1,M_2,M_3) = \frac{3!\,\Gamma(3)}{\Gamma(6)} = \frac{1}{10}$$

Supplemental Exercise 13.5.4: Place your finger on the irony of Feller's explanation for Bose–Einstein statistics.

Solution to Supplemental Exercise 13.5.4:

The Bose–Einstein statistics fall out quite naturally by adopting Laplace's uniform prior for one of the terms demanded by the formal rules of probability theory, the very same prior that must be used in order to derive his **Rule of Succession**! I want to be sure that the delicious irony of this outcome has not escaped your notice. Feller's scathing contempt of Laplace when Laplace used his **Rule of Succession** to find the probability that the Sun would rise tomorrow is well known.

And yet, in the final analysis, it is Laplace's correct employment of the rules of probability that provide the only non–mysterious rationale for these so–called Bose–Einstein statistics. Einstein's and Feller's, or the ensuing universal invocation of "indistinguishability" of whatever is supposed distributed over the statements in the state space, never furnished, at least to my way of thinking, an equally compelling rationale as Laplace had provided.

Moreover, probability theory would call the left hand side of the equation in the last exercise something very conventional like, "The probability of future frequency counts in the state space when no data were observed." It would not think to employ such bizarre language coming out of left field that talked of "Bose–Einstein" statistics. How could an IP ever place every single occurrence from the state space into the appropriate cell of the contingency table if that occurrence did not possess some "distinguishable" label like a name, a color, a temporal ordering, or some other "distinguishing" feature?

Supplemental Exercise 13.5.5: Construct one final example of Bose–Einstein statistics.

Solution to Supplemental Exercise 13.5.5:

Let's take one final look at a detailed numerical example of a sample space and the accompanying Bose–Einstein statistics. This exercise treats an example that is slightly larger than Feller's $r = 3$ together with $n = 3$ or $n = 5$ examples that have already been closely scrutinized. Suppose we select $r = 5$ and $n = 8$ to represent 5 distinguishable balls, **a, b, c, d, e**, distributed over $n = 8$ cells.

One familiar inferential scenario is the kangaroo state space where each kangaroo possesses three binary traits. See Chapter Twenty Two, Volume II, and especially Figure 22.1, where the set up of the joint probability table shows the indexing of all eight joint statements.

Thus, we will be looking at $M = 5$ kangaroos distributed over $n = 8$ cells of a contingency table. Cell 1 indexes the joint occurrence of being right handed, Foster's

drinking, and having beige fur color, and, in like manner all the way through to cell 8 indexing the joint occurrence of being left handed, Corona drinking, and having sandy fur color.

There are $n^r = n^M = 8^5 = 32,768$ elementary points in this sample space. Any one of these elementary points is the most detailed event we can describe. Such an event might be that two kangaroos named Bruno and Dino are left handed sandy colored Foster's drinkers, two kangaroos named Arlo and Elmo are right handed beige colored Corona drinkers, and the final kangaroo named Carlo is a left handed beige colored Foster drinker. Feller's description would have the balls **b, d** placed in cell 4, the two balls **a, e** in cell 5, while the final ball **c** is in cell 2.

This is merely one of the 30 possibilities that exist for the situation where one kangaroo is in cell 2, two kangaroos are in cell 4, and the final two kangaroos are in cell 5. For example, another possibility from the 30 available would have Arlo and Bruno in cell 4, Carlo and Dino in cell 5, and Elmo in cell 2.

Moving up one more level, there are 168 examples of contingency tables like this one just presented that exhibit the property that two kangaroos possess one set of traits, two more kangaroos share a different set of traits, with the final kangaroo possessing its own distinct traits different from the other four kangaroos.

In other words, focus on the particular integer partition of the sum of the future frequency counts $M = 5$ as $2 + 2 + 1 + 0 + 0 + 0 + 0 + 0 = 5$. There are 168 ways for this to occur. One of these 168 possibilities for the integer partition $2+2+1+0+0+0+0+0$ is the contingency table $\boxed{0}\boxed{1}\boxed{0}\boxed{2}\boxed{2}\boxed{0}\boxed{0}\boxed{0}$ just used as the example.

Table 13.6 on the next page summarizes all seven occupancy patterns in the same format as before. The third column labeled *First* applies the first counting formula to find the contribution to the total number of contingency tables,

$$\textbf{Partial sum of contingency tables} = \frac{n!}{r_z!\, r_s! \cdots r_M!}$$

The fourth column, labeled *Second*, applies the second counting formula, that is, the multiplicity factor, for the number of ways each such contingency table might arise when taking account the distinguishability of the kangaroos,

$$\textbf{Multiplicity factor} = \frac{M!}{M_1!\, M_2! \cdots M_8!}$$

The sum over column three is the total number of contingency tables,

$$\textbf{Total number of contingency tables} = \frac{(M+n-1)!}{M!\,(n-1)!} = \frac{(8+5-1)!}{5!\,7!} = 792$$

The sum over the final column is the total number of elementary points in the sample, $n^M = 8^5 = 32,768$.

Table 13.6: *Accounting for all of the elementary points in the sample space for $M = 5$ kangaroos and $n = 8$ traits.*

Pattern	*Integer Partition*	*First*	*Second*	*Elementary points*
1	$5+0+0+0+0+0+0+0$	8	1	8
2	$4+1+0+0+0+0+0+0$	56	5	280
3	$3+2+0+0+0+0+0+0$	56	10	560
4	$3+1+1+0+0+0+0+0$	168	20	3360
5	$2+2+1+0+0+0+0+0$	168	30	5040
6	$2+1+1+1+0+0+0+0$	280	60	16800
7	$1+1+1+1+1+0+0+0$	56	120	6720
Totals		792		32768

We have been discussing Occupancy Pattern #5. The number of contingency tables contributed to the overall total is found by,

$$\textbf{Partial sum of contingency tables} = \frac{n!}{r_z!\, r_s!\, \cdots\, r_M!} = \frac{8!}{5!\, 1!\, 2!\, 0!\, \cdots\, 0!} = 168$$

as shown under column *First*. The multiplicity factor for these contingency tables is,

$$W(M) = \frac{M!}{M_1!\, M_2\, \cdots\, M_n!} = \frac{5!}{0!\, 1!\, 0!\, 2!\, 2!\, 0!\, 0!\, 0!} = 30$$

as shown under column *Second*. The contribution to the total number of elementary points for this occupancy pattern is then $168 \times 30 = 5040$ as shown under the final column *Elementary points*.

Now that the detailed development for all of the numbers in this inferential scenario are out front and center, we can move on to the primary issue of probability assignments. As Feller says, one could continue to treat this problem just like his introductory example of sample spaces, elementary points, and events. In this case, each elementary point has a probability of 1/32768.

Consider the compound event **Event A** where only two kangaroos share a trait, while the other three all have different traits. This is the compound event defined by Occupancy Pattern #6, the integer partition of $M = 5$ by $2+1+1+1+0+0+0+0$. There are 16,800 elementary points comprising this event, so according to Feller if one maintains distinguishability of the kangaroos, *the* probability *of* **Event A** is,

$$P(\textbf{Event A}) = \frac{16,800}{32,768} = 0.5127$$

Of course, to invoke Bose–Einstein statistics, Feller asserts without any justifying rationale from probability theory, that now the kangaroos are *indistinguishable*! The

new probability assignment to **Event A** becomes,

$$P(\textbf{Event A}) = \frac{280}{A_{r,n}} = \frac{280}{\binom{n+r-1}{r}} = \frac{280}{\binom{12}{5}} = \frac{280}{792}$$

The rules of probability theory happen to provide a consistent rationale for both situations. Therefore, why invoke unnecessary and dubious new principles like *indistinguishability*? The guiding template is simply the formal rule,

$$P(M_1, M_2, \cdots, M_n) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, \cdots, M_n \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)$$

In the first case, *only one specific model* $\mathcal{M}_k$ appears in the above equation. Each elementary point has a probability assignment of 1/32768, and the probability of compound events is an aggregation formed by summing over these elementary points. In fact, that one model is the assignment of $Q_i = 1/n$, which I have been calling the "fair model." If that happens to be the only model appearing in the summation, the technical consequence of employing the Dirac delta function for $P(\mathcal{M}_k)$ is the multinomial distribution. The α_i parameters in pdf (q) are equal and march in lockstep to ∞ leading to $\delta(q - 1/n)$,

$$P(M_1, M_2, \cdots, M_n) = W(M)\; Q_1^{M_1}\; Q_2^{M_2}\; \cdots\; Q_n^{M_n}$$

However, still not straying one inch from the strict confines of the formal rules, and, furthermore, not dredging up any new labels such as "Bose–Einstein statistics" to add to the confusion, the second situation is also fully covered by the guiding template. Because the IP does not wish to claim that much knowledge about the models making the probability assignments as in the first situation, like Laplace, the IP claims that it is totally uninformed about the models, and duly makes $P(\mathcal{M}_k)$ not a Dirac delta function, but rather a Dirichlet distribution with all α_i parameters equal to 1. The technical consequence from adopting this stance is that the probability for any future frequency counts becomes,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!\;(n-1)!}{(M+n-1)!} = \frac{1}{792}$$

the same as Feller's "Bose–Einstein statistics."

Table 13.7 lists the probabilities of seven compound events under Feller's criteria for "Bose–Einstein statistics" and "Maxwell–Boltzmann statistics." The events listed under the first column are the same those defined by the occupancy pattern numbers. Thus, row 4 is the integer partition $3 + 1 + 1 + 0 + 0 + 0 + 0 + 0 = 5$.

This is the event where three kangaroos share the same trait, one kangaroo has a different trait than these three, and the final kangaroo has yet a different trait from the preceeding kangaroos. The contingency table $\boxed{0}\boxed{0}\boxed{3}\boxed{0}\boxed{0}\boxed{0}\boxed{1}\boxed{1}$ is an example where three kangaroos share the trait of sandy fur colored right handed Foster's drinkers, one kangaroo is a sandy fur colored right handed Foster's drinker, while the fifth and final kangaroo is a sandy fur colored left handed Corona drinker.

The probability for Event 4 under "Bose–Einstein statistics" is,

$$P(\textbf{Event 4}) = 168/792$$

while the probability under "Maxwell–Boltzmann statistics" is,

$$P(\textbf{Event 4}) = 3360/32768$$

Table 13.7: *Comparison of the probabilities for events under Maxwell–Boltzmann and Bose–Einstein statistics.*

Event	*Description*	*Bose–Einstein*	*Maxwell–Boltzmann*
1	five, none	8/792	8/32768
2	four, one	56/792	280/32768
3	three, two	56/792	560/32768
4	three, one, one	168/792	3360/32768
5	two, two, one	168/792	5040/32768
6	two, one, one, one	280/792	16800/32768
7	none, five	56/792	6720/32768

Compare the ramifications for the probabilities of events under an IP's two differing states of knowledge about the causes of events. Under what Feller labels "Maxwell–Boltzmann statistics," and what I call the adoption of the single fair model, the probability of observing the event where all five kangaroos share the same trait is quite low at 8/32786. Under what Feller labels as "Bose–Einstein statistics," and what I call an adoption of a uniform prior probability over *all* models, the probability of this event is much higher at 8/792.

The most probable event under either state of knowledge about the models is in row 6 where only two kangaroos share a trait. The remaining three kangaroos all possess different traits. But even here, the probability under "Maxwell–Boltzmann statistics" is a little over 50% at 16800/32786, when compared to the much smaller probability of only 280/792, about 35%, under "Bose–Einstein statistics."

It is clear that "extreme" events like all M kangaroos sharing the same trait has a higher probability when an IP admits to "total ignorance," as opposed to when it claims far more knowledge about causes when *one single model* is selected.

It is quite fascinating that these different values for the probabilities can be interpreted in one manner, as Feller chose to do, by characterizing kangaroos as suddenly "indistinguishable," when heretofore the mere fact that they had to be placed into the appropriate cells of the contingency table meant that they must in fact be "distinguishable." Contrary to Feller, I find it more appealing to rephrase things such that the fundamental curiosity revolves around the complementary roles of the multiplicity factor and the α_i parameters of the Dirichlet distribution.

As all the $\alpha_i \to \infty$ at the same rate, then the one single model, the fair model, is recovered. The multiplicity factor reigns supreme. When the $\alpha_i = 1$, we are back to Laplace's invocation of an IP's total ignorance about causes. The multiplicity factor doesn't count for anything. When the $\alpha_i \to 0$, ignorance on the part of the IP is moved to a different place, while the role of the multiplicity factor is completely reversed! Those events where all M kangaroos share the same trait approach certainty. Those events with a multiplicity factor of 1 possess the *largest* probability. Those events that can happen in the greatest number of ways have an insignificant probability.

All of these curious happenings are easily recognized only if the IP maintains a strict discipline with regard to the formal rules of probability theory. There is no need to indulge in any kind of creative invocation of novel concepts or creative labeling outside the purview of the axiomatic foundations.

Feller's mandatory indistinguishable entities, whether they be kangaroos, coins, dice, students, or photons, do not appear as part of the foundational concepts. Moreover, they may, in fact, be indistinguishable to some IP, but they still must be correctly placed into the contingency table in repeated trials. They may be indistinguishable to some IP, but not to Nature herself.

The concept of indistinguishability fails as a justifiable rationale. The rationale that an IP possesses differing states of knowledge about the causes of events is, however, part of the foundational concepts. That is all that is required to produce the probabilities appearing under the unnecessary rubric of "Maxwell–Boltzmann," "Bose-Einstein," or "Fermi–Dirac."

Supplemental Exercise 13.5.6: Discuss a new example of Feller's third unnecessary label, "Fermi–Dirac statistics."

Solution to Supplemental Exercise 13.5.6:

Feller presented his numerical example of Fermi–Dirac statistics [6, pg. 42, Vol 1] with $r = 3$ and $n = 5$. I examined Feller's example in detail in Exercise 27.7.18 of Volume II.

For another numerical example of Fermi–Dirac statistics, pick different numbers of balls and cells. Suppose that $r = 7$ and $n = 10$. We have by a law of Physics, the Pauli exclusion principle, that, for example, no two electrons can occupy the same quantum state. To meet the requirement that no two or more balls can co–exist in the same cell, r must be less than or equal to n which our example satisfies.

For Fermi–Dirac statistics, the number of integer partitions collapses down to the simplest imaginable. There is always only one integer partition. Here it is,

$$M = \overbrace{1+1+1+1+1+1+1}^{7} + \overbrace{0+0+0}^{3} = 7$$

The total number of contingency tables simplifies as well since it is just the partial sum for the only existing pattern for the contingency table as just written down,

$$\textbf{Partial sum of contingency tables} = \frac{n!}{r_z!\, r_s!\, \cdots\, r_M!} = \frac{n!}{r_z!\, r_s!} = \binom{n}{r_z}$$

since there can only be repetitions of the zero count and the single count.

Feller gives the probability under Fermi–Dirac statistics as $\binom{n}{r}^{-1}$. So both Feller and I agree that for this numerical example, the probability of an event like the contingency table $\boxed{1}\boxed{0}\boxed{1}\boxed{1}\boxed{1}\boxed{0}\boxed{1}\boxed{1}\boxed{0}\boxed{1}$ is,

$$P(\textbf{Event}) = \frac{1}{\frac{10!}{7!\,3!}} = \frac{1}{120}$$

I claimed that no special rationale such as Feller's "specially labeled statistics" was required to reproduce any probability of an event. All that *was* required was the discipline to continue to follow the formal rules for probability manipulations. Here, the formal rule that is our guiding template is the probability for the future frequency counts as a marginal probability over the counts conditioned on all models times the prior probability of the models,

$$P(M_1, M_2, \cdots, M_n) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, \cdots, M_n \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

We have already tested the truth of this claim for Maxwell–Boltzmann and Bose–Einstein statistics. To reproduce Feller's Fermi–Dirac statistics, all we have to do is continue to rely upon the formula derived from the guiding template,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})} \times \prod_{i=1}^{n} \frac{\Gamma(M_i+\alpha_i)}{M_i!\,\Gamma(\alpha_i)}$$

Because the restriction has been levied on us for this problem that all the M_i must be 0 or 1, the formula collapses to,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})}$$

The right most term,

$$\prod_{i=1}^{n} \frac{\Gamma(M_i+\alpha_i)}{M_i!\,\Gamma(\alpha_i)}$$

will reduce to 1 because the M_i are all either 0 or 1, and more significantly, because the α_i parameters are set to 1 and approaching 0.

If the parameters in the Dirichlet distribution for the prior probability of the models are set such that,

$$\sum_{i=1}^{n} \alpha_i = \mathcal{A} = n - M + 1$$

then,

$$P(M_1, M_2, \cdots, M_n) = \frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})} = \frac{M!\,(n-M)!}{n!} = \frac{r!\,(n-r)!}{n!} = \binom{n}{r}^{-1}$$

For our current example,

$$\sum_{i=1}^{10} \alpha_i = \mathcal{A} \equiv n - M + 1 = 10 - 7 + 1 = 4$$

so that any four α_i parameters are set to 1, and the remaining six are set to approach 0. Any event where $\sum_{i=1}^{n} M_i = M \equiv r = 7$ the M_i are either 0 or 1, and $n = 10$,

$$P(M_1, M_2, \cdots, M_{10}) = \frac{7!\,3!}{10!} = \frac{1}{120}$$

It doesn't make any difference which four $\alpha_i = 1$, and which six $\alpha_i \to 0$. Any one of the 120 possible contingency tables will then have the same probability.

This set up of the integer partition of M for Fermi–Dirac statistics and the Pauli exclusion principle is similar to that of the *Birthday Problem*. There we also set up an integer partition where, say, 30 people all had a different birthday over 365 days of the year.

The moral of these exercises is that we have been able to reproduce Feller's "Maxwell–Boltzmann," "Bose–Einstein," and "Fermi–Dirac" statistics simply by following the formal manipulation rules of probability. This goal was achieved through the freedom to change the information in the prior probability over model space. Examining different configurations of the α_i parameters in the Dirichlet probability density function was the same as inserting alternative information into the prior probability $P(\mathcal{M}_k)$, and allowed an IP to reconstruct Feller's "statistics."

13.6 Back to Predicting Success in College

After all of these fascinating diversions, let's return to the main real world inferential problem of Volume I's Chapter Thirteen. Can an IP leverage probability theory to generalize logic, and predict college graduation success for any number of potential students after they have taken some aptitude tests? Chapter Thirteen took the tack of first examining fundamental conceptual issues in probability theory that must be confronted before dealing with any data.

I wanted to draw a clear distinction between the orthodox view based on Feller's sample space and my presentation emphasizing the model space. It is true that much of our preliminary work has focused on various counting formulas which retain their legitimacy under either viewpoint. But I had hoped to delineate a rather sharp conceptual distinction between these combinatorial numbers and the numbers representing a degree of belief in the truth of some joint statement about a student's graduation outcome together with his or her test results.

Suppose for our subsequent numerical exercises, we up the ante to a state space with dimension $n = 8$. Instead of just one test, the students are scheduled to take two tests with each test having a binary measurement outcome of HIGH or LOW.

As before, the graduation statement is also binary with measurement outcomes of GRADUATES or DOES NOT GRADUATE. By definition, these are the only mutually exclusive and exhaustive observations that can be made by the IP. With this complement of three binary statements A, B, and C, the state space, as well as any joint probability table, or any contingency table will be defined with an $n = 8$.

There are six students who will take the two tests. They are distinguishable and they have names: (1) Ariel, (2) Barrie, (3) Carlos, (4) Devinder, (5) Edwin, and (6) Frančois. When we take up Feller's sample space, the abstraction dictates that six balls, **a, b, c, d, e**, and **f** are thought to be distributed over eight cells. For convenience, these cells are arranged in a linear fashion.

The total number of elementary points in the sample space will be calculated as $n^r \equiv n^M = 8^6 = 262,144$. There will be a total of,

$$\frac{(M+n-1)!}{M!\,(n-1)!} = \frac{13!}{6!\,7!} = 1,716$$

possible contingency tables reflecting any future data, namely any future frequency counts M_1 through M_8. These $1,716$ contingency tables are broken down into eleven categories as shown in Table 13.8 according to the number of students sharing a trait. This is the integer partition for $M = 6$ over $n = 8$ cells.

Table 13.8: *Accounting for all of the elementary points in the sample space for $M = 6$ students and a state space for graduation and two aptitude tests of $n = 8$.*

Category	*Integer Partition*	*First*	*Second*	*Elementary points*
1	$6+0+0+0+0+0+0+0$	8	1	8
2	$5+1+0+0+0+0+0+0$	56	6	336
3	$4+2+0+0+0+0+0+0$	56	15	840
4	$4+1+1+0+0+0+0+0$	168	30	5040
5	$3+3+0+0+0+0+0+0$	28	20	560
6	$3+2+1+0+0+0+0+0$	336	60	20160
7	$3+1+1+1+0+0+0+0$	280	120	33600
8	$2+2+2+0+0+0+0+0$	56	90	5040
9	$2+2+1+1+0+0+0+0$	420	180	75600
10	$2+1+1+1+1+0+0+0$	280	360	100800
11	$1+1+1+1+1+1+0+0$	28	720	20160
Totals		1716		262144

The first counting term involved in the multiplication shown in the final column finds the number of possible contingency tables for any one of the eleven occupancy patterns. For example,

$$\frac{n!}{r_z!\, r_s! \cdots r_M!} = \frac{8!}{6!\ 1!\ 0!\ 0!\ 0!\ 1!\ 0!} = 56$$

tells us that there are 56 possible future frequency counts where there exist six repetitions of a zero count, one repetition of a single count, and one repetition of a five count. This is the occupancy pattern shown in the second row. The second counting term is calculated by the multiplicity factor $W(M)$ for any one these 56 contingency tables,

$$W(M) = \frac{M!}{M_1!\, M_2!, \cdots, M_8!} = \frac{6!}{0!\, 5!\, 1!\, 0!\, 0!\, 0!\, 0!\, 0!} = 6$$

This multiplicity factor happens to be attached to $\boxed{0}\boxed{5}\boxed{1}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}$ indicating that there are six different ways that one of the six named students could have been the singleton count in cell 3, the student who graduates with a HIGH score on the first test and a LOW score on the second test while still graduating. Therefore, this occupancy pattern contributes 336 elementary points to the overall total of $262,144$ elementary points in this sample space.

One such elementary point from the 336 is that Barrie GRADUATED, but was the only student who scored HIGH on the first test and LOW on the second, while the other five students Ariel, Carlos, Devinder, Edwin, and Frančois also GRADUATED, but performed the opposite of Barrie, scoring LOW on the first test and HIGH on the second test. The other five possibilities indicated by the multiplicity factor are that Ariel was the one student who had the same traits as just detailed for Barrie, that Carlos was the one student ..., and so on.

This same multiplicity factor is going to add the same number of elementary points to the total for the contingency table $\boxed{0}\boxed{0}\boxed{0}\boxed{1}\boxed{0}\boxed{0}\boxed{5}\boxed{0}$, which is just another example of the occupancy pattern of the second row where five students share the same trait while the remaining student is different. Here, five students score HIGH on the first test, LOW on the second test, but DO NOT GRADUATE. The remaining student, whoever he is, scores LOW on both tests, but still manages to GRADUATE.

Notice that the occupancy pattern in the last row where all six students possess different traits has the largest multiplicity factor of any occupancy pattern at 720. However, that doesn't mean that it contributes the most number of elementary points to the total number in the sample space. As a consequence, the most spread out distribution of students is not going to be the most probable occupancy pattern.

Supplemental Exercise 13.6.1: According to Feller's sample space view of probability, which occupancy pattern is most probable?

Solution to Supplemental Exercise 13.6.1:

This would be the occupancy pattern in row 10 with two students sharing the same traits, and the remaining four students possessing different traits. Since this pattern contains 100,800 elementary points, *the* probability *of* this event is,

$$P(\textbf{Event}) = \frac{100,800}{8^6} = \frac{100,800}{262,144} = 0.38$$

The second most probable occupancy pattern is in row 9 where two sets of two students share the same trait, while the remaining three students are different from both sets. Together, the event compounded from row 9 and row 10 has *the* probability *of*,

$$P(\textbf{Compound Event}) = \frac{100,800 + 75,600}{8^6} = \frac{176,400}{262,144} = 0.67$$

Supplemental Exercise 13.6.2: Suppose, contrary to Feller's viewpoint, that the information in some ECA Rule provides the model for assigning probabilities. What now is the probability for an event?

Solution to Supplemental Exercise 13.6.2:

Feller's definition of a probability from the sample space approach asserts, in the final analysis, that the IP adopts just one model for assigning numerical values to the probabilities of the joint statements in the state space. If the IP is restricted to just one model, then it is prevented from entertaining what it considers to be more plausible models than the assignment provided under the "fair model."

One example of such a plausible model as entertained by an IP might be where it is considered impossible for a student to graduate if they score LOW on both tests, or likewise, impossible not to graduate if they score HIGH on both tests.

Such a model would dictate an assignment of 0 in cells 4 and 5 of the joint probability table. In addition, this plausible model asserts that the graduation rate is higher than not graduating, so the marginal probability for GRADUATES must reflect this fact. This model then assigns a probability of 1/4 to cells 1, 2, and 3, with the aforementioned 0 already placed in cell 4 for a marginal probability of $P(\text{GRADUATES}) = 3/4$. The assignment to cells 6, 7, and 8 is then 1/12 with the other aforementioned 0 already placed into cell 5 by the model.

This is a model inspired by ECA Rule 231. The functional assignments to the Boolean function with three arguments is $f(T, F, F) = F$ and $f(F, T, T) = F$ with the other six possibilities having a functional assignment of T. Examine the typical deciphering for Wolfram's rule number in Table 13.9.

Table 13.9: *Translating a model into an ECA Rule Number. The second row shows the labeling for graduation and test outcomes.*

TTT	TTF	TFT	TFF	FTT	FTF	FFT	FFF
GHH	GHL	GLH	GLL	$NGHH$	$NGHL$	$NGLH$	$NGLL$
T	T	T	F	F	T	T	T
1	1	1	0	0	1	1	1
2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
128	64	32	0	0	4	2	1
Rule number 231							

In order not to unduly tax our brains, and to illustrate the principles involved, look at the first occupancy pattern in row 1 of Table 13.8. All six students share the same trait. There are obviously only eight possible contingency tables. But two of these contingency tables are excluded as impossible by the model. These future frequency counts excluded by the model are where all six students occur in cell 4, and where all six students occur in cell 5.

For the three remaining possible contingency tables covering GRADUATES, we have the probability,

$$3 \times W(M) \times Q_1^{M_1} Q_2^{M_2} \cdots Q_8^{M_8} = 3 \times 1 \times (1/4)^6 = 0.000732$$

For the three remaining possible contingency tables for DOES NOT GRADUATE, we have the probability,

$$3 \times W(M) \times Q_1^{M_1} Q_2^{M_2} \cdots Q_8^{M_8} = 3 \times 1 \times (1/12)^6 = 0.000001$$

For this event then, we have the probability relying upon the information from the one model inspired by ECA Rule 231,

$$P(\textbf{Event}) = 0.000732 + 0.000001 = 0.000733 = 7.33 \times 10^{-4}$$

Compare this probability to Feller's probability of,

$$P(\textbf{Event}) = \frac{8}{8^6} = 0.000031 = 3.05 \times 10^{-5}$$

Finally, compare both of these probabilities based on the information in just one model to our preferred probability based on considering all possible models equally,

$$P(\textbf{Event}) = \frac{8}{1716} = 0.00466 = 4.66 \times 10^{-3}$$

Supplemental Exercise 13.6.3: What are some of the complications which ensue when we consider models other than the fair model?

Solution to Supplemental Exercise 13.6.3:

In the last exercise, we have seen that two possible future frequency counts were impossible under the information from a model inspired by Rule 231. Things get stickier when calculating the probabilities for other events.

Shift your attention to an event defined by two sets of three students sharing the same trait. This is the integer partition, or as Feller calls it, the occupancy number, listed in row 5 of Table 13.8. There are 28 possible future frequency counts for this situation.

But, somewhat surprisingly, almost half of these, thirteen contingency tables to be exact, are ruled out as impossible under this new alternative inspired by Rule 231. A future frequency count of 3 in either cell 4 or cell 5 would be grounds for elimination.

For example, one of the 28 possible contingency tables is $\boxed{0}\boxed{0}\boxed{0}\boxed{3}\boxed{0}\boxed{3}\boxed{0}\boxed{0}$, indicating three students who GRADUATE with LOW scores on both tests, and three students who DO NOT GRADUATE with a LOW score on the first test and a HIGH score on the second test.

However, this future contingency table is impossible under our specified model since the probability assigned to cell 4 indexing students who GRADUATE with LOW scores on both tests is 0. The future contingency tables would have to be checked to see which are admissible and which are impossible. Furthermore, the probability for the admissible contingency tables would change depending on whether both sets of three students GRADUATE, DO NOT GRADUATE, or are split between the two.

For example, one of these fifteen admissible contingency tables, or alternatively, future frequency counts, is $\boxed{3}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{3}\boxed{0}$. We see that one set of three students GRADUATES with HIGH scores on both tests. The second set of three students DOES NOT GRADUATE with a HIGH score on the first test and a LOW score on the second test.

Once again, every admissible contingency table would have to be checked to see which category it falls into. Of the fifteen admissible tables, there are nine tables, like the example just presented, where the three students are split between GRADUATE and DOES NOT GRADUATE. There are three tables where both sets of three students GRADUATE, and three tables where both sets of three students DO NOT GRADUATE.

The probability contributed to the probability for the overall event by the nine

tables is,

$$9 \times W(M) \times Q_1^{M_1}\, Q_2^{M_2} \cdots Q_8^{M_8} = 9 \times \frac{6!}{3!\,3!} \times (1/4)^3 \times (1/12)^3 = 0.00163$$

The probability contributed to the probability for the overall event by the three tables where both sets of students GRADUATE is,

$$3 \times W(M) \times Q_1^{M_1}\, Q_2^{M_2} \cdots Q_8^{M_8} = 3 \times \frac{6!}{3!\,3!} \times (1/4)^6 = 0.01465$$

The probability contributed to the probability for the overall event by the three tables where both sets of students DO NOT GRADUATE is,

$$3 \times W(M) \times Q_1^{M_1}\, Q_2^{M_2} \cdots Q_8^{M_8} = 3 \times \frac{6!}{3!\,3!} \times (1/12)^6 = 0.00002$$

leading to the probability of the event,

$$P(\textbf{Event}) = 0.01630$$

Once again, compare this probability to Feller's probability of,

$$P(\textbf{Event}) = \frac{560}{262,144} = 0.00214$$

Finally, compare both of these probabilities both based on the information in just one model to our preferred probability based on considering all possible models,

$$P(\textbf{Event}) = \frac{28}{1716} = 0.01632$$

We observe a curious numerical coincidence where the probability of two sets of three students sharing the same trait is almost the same when the single model from Rule 231 is used, or when all conceivable models with no one model having higher probability than another are used.

Supplemental Exercise 13.6.4: What is the probability that a student GRADUATES given that he or she scored HIGH on the first test and LOW on the second test?

Solution to Supplemental Exercise 13.6.4:

This solution demands an application of Bayes's Theorem. The generic formula looks like,

$$P(A \,|\, B, C) = \frac{P(A, B, C)}{P(B, C)}$$

and for our particular application here,

$$P(\text{Status} = \text{GRADUATES} \,|\, \text{Test1} = \text{HIGH}, \text{Test2} = \text{LOW}) = \frac{P(G, H, L)}{P(H, L)}$$

$$= \frac{P(G, H, L)}{P(G, H, L) + P(\overline{G}, H, L)}$$

Now, since these probability expressions do not show the presence of any models, they must have been marginalized over,

$$P(G, H, L) = \sum_{k=1}^{\mathcal{M}} P(G, H, L \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

If an uninformed IP follows Laplace's advice, this becomes,

$$P(G, H, L) = \frac{M!\,(n-1)!}{(M+n-1)!}$$

Here, the IP is asking for just the very next occurrence of a particular joint statement. Since there are no data, there are no previous occurrences, and the very next occurrence happens to be the very first occurrence. Thus, $M = 1$, with $M_3 = 1$ and the remaining $M_i = 0$,

$$\begin{aligned} P(G, H, L) &= \frac{M!\,(n-1)!}{(M+n-1)!} \\ &= \frac{1!\,7!}{8!} \\ &= \frac{1}{8} \\ &= P(M_1 = 0, M_2 = 0, M_3 = 1, \cdots, M_8 = 0) \end{aligned}$$

The number we have arrived at, $1/n$, is effectively Feller's prescription for the assigned probability to a joint statement.

We have been exposed to this numerical coincidence before. The probability of HEADS on a coin toss is $1/2$ under the fair model; the probability for a THREE on a roll of a die is $1/6$ under the fair model; the probability for any one of the eight joint statements in our current example is $1/8$ under the fair model. And these probabilities for whatever is under consideration are exactly the same as found for the very first occurrence when averaged over all models having an equal prior probability.

Similarly, the second probability required in the denominator of Bayes's Theorem is calculated as,

$$P(\overline{G}, H, L) = \sum_{k=1}^{\mathcal{M}} P(\overline{G}, H, L \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

If an uninformed IP follows Laplace's advice, this becomes,

$$\begin{aligned} P(\overline{G}, H, L) &= \frac{M!\,(n-1)!}{(M+n-1)!} \\ &= \frac{1!\,7!}{8!} \\ &= \frac{1}{8} \\ &= P(M_1 = 0, M_2 = 0, \cdots, M_7 = 1, M_8 = 0) \end{aligned}$$

Substituting into Bayes's Theorem, we find that the probability for a student to GRADUATE given that he or she *MIGHT* (the subjunctive mood) score HIGH on the first test and LOW on the second test is 1/2,

$$\begin{aligned} P(\text{Status} = \text{GRADUATES} \,|\, \text{Test1} = \text{HIGH}, \text{Test2} = \text{LOW}) &= \frac{P(G, H, L)}{P(G, H, L) + P(\overline{G}, H, L)} \\ &= \frac{1/8}{1/8 + 1/8} \\ &= \frac{1}{2} \end{aligned}$$

As we have discussed many times before, the difference between adopting a single model and averaging all models only becomes apparent when we inquire about the next *two* occurrences, or the next *three* occurrences, ..., or the next M occurrences.

If the IP adopts the single fair model, then the probability for the next two students to GRADUATE with the above test scores is,

$$P(\textbf{Event}) = W(M) \times Q_1^{M_1} \cdots Q_8^{M_8} = 2 \times Q_3^2 = 2 \times (1/8)^2 = 1/32$$

On the other hand, if the IP is uninformed about the causal mechanism behind test scores and graduation status, then the probability is, (Refer back to Supplemental Exercise 13.5.3),

$$\begin{aligned} P(\textbf{Event}) &= \frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})} \times \prod_{i=1}^{8} \frac{\Gamma(M_i+1)}{M_i!\,\Gamma(1)} \\ &= \frac{2!\,\Gamma(8)}{\Gamma(10)} \times \frac{\Gamma(3)}{2!\,\Gamma(1)} \\ &= \frac{1}{36} \end{aligned}$$

Check this answer for coherency. Everything that could possibly happen for two future frequency counts must have a probability of 1. We have just examined the case where two students had a particular trait graduating with a high score on test 1 and a low score on test 2. This is the one contingency table $\boxed{0}\boxed{0}\boxed{2}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}$. There are seven other contingency tables just like this with two students in a cell and no students in the remaining seven cells. They all have the same probability of 1/36. This accounts for 8/36 of the certainty.

But to account for everything that could happen for two future frequency counts, there are some number of contingency tables covering the case where one student has one trait and the other student has a different trait. Using our familiar formula for the number of such possible contingency tables we find that there are 28 in total. We could calculate the probability for each one of these contingency tables just like we did above, but we know from conceptual principles governing the uninformed situation that the probability for every one of these 28 contingency tables must also equal 1/36. Coherency is maintained; the probability is calculated as

$$\frac{8}{36} + \frac{28}{36} = 1$$

for everything that could happen.

Remember the canonical 99 future coin flips for an uninformed IP where we find that the probability of no HEADS is 1/100, the probability for 50 HEADS is 1/100, the probability for 99 HEADS is 1/100? Laplace's conceptual principle applies here as well for any type of future frequency counts; they all have a probability of,

$$\frac{M!\,(n-1)!}{(M+n-1)!} = 1/36$$

Once again, let me remark on the fact that the multiplicity factor $W(M)$ plays a significant role if the IP adopts a single model, whereas it plays no role whatsoever if the IP averages over all models that have equal prior probability.

Chapter 14

Predicting College Success When Data Are Available

14.1 The No Data Calculations

The previous Chapter's work was mainly a matter of principle. It has very little pragmatic significance. While admitting to the correctness of the fundamental probabilistic principles as they are applied to the no data case, a practicing scientist will always choose to make inferences on empirical data and the best scientific models.

But, as Jaynes was wont to say, these exercises served as a pump priming for the situation we really want to study. These fundamental principles of probability carry over very nicely when an IP wants to condition on the presence of known observations.

The most salient change when dealing with data is the opportunity for the IP to update an initial uninformed model space to one where models do have different probabilities. This updating takes place through the auspices of Bayes's Theorem to generate the posterior probability for models. This updating procedure is written generically as,

$$
\begin{aligned}
P(\mathcal{M}_k \,|\, \mathcal{D}) &= \frac{P(\mathcal{D}, \mathcal{M}_k)}{P(\mathcal{D})} \\
&= \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}{P(\mathcal{D})} \\
&= \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)}
\end{aligned}
$$

Before plunging into the calculations for the probability of any future frequency counts conditioned on the available data, let's review the slightly changed inferential scenario involving college students and aptitude tests as it has been enhanced from the last Chapter. By incorporating a third aptitude test with a binary outcome, the dimension of the state space has been increased from $n = 8$ to $n = 16$. Any joint probability table, or contingency table, involved in this scenario will consist of 16 cells. Each cell indexes some joint statement involving a student's graduation status, and scores on all three tests.

The M and M_i notation referring to future frequency counts remains the same. Since there is now data available, we bring in the N and N_i notation referring to any past frequency counts already recorded in the contingency table. The multiplicity factor for the future data is $W(M)$, and for the past data $W(N)$.

The α_i always refer back to the parameters of the Dirichlet distribution. Most of the time, these $\alpha_i = 1$ in order to implement a uniform distribution over any possible assignment to the probabilities of the joint statements in the state space. The index reflected in the subscript i always runs from $i = 1$ to $i = n$.

Supplemental Exercise 14.1.1: Provide another example of one of the sixteen possible joint statements involving the college student graduation scenario with three tests.

Solution to Supplemental Exercise 14.1.1:

Focus on cell 13 of the contingency table in Volume I's Figure 14.1. This cell indexes the occurrence of the joint statement, "The student does NOT GRADUATE after having scored LOW on test B, HIGH on test C, and HIGH on test D."

This statement is a potential occurrence in the real world, and can be judged by an IP as TRUE or FALSE. If it is TRUE, then the frequency count in cell 13 of the contingency table is increased by one. If it is FALSE, then one of the other fifteen possible occurrences from the total state space of $n = 16$ must have been TRUE for that particular observation.

The probability, or degree of belief, held by the IP that the above joint statement is TRUE, might be expressed symbolically as,

$$P(A = a_2, B = b_2, C = c_1, D = d_1) \text{ or as } 0 \leq P(X = x_{13}) \leq 1$$

It is important to note that these above expressions are what I have labeled as *abstract probabilities*. In other words, the expressions for the probability show no conditioning on some given model. In all cases, we should in fact disambiguate any abstract probability by indicating the information entering into the probability distribution through model $\mathcal{M}_k$ with the symbolic expression $P(X = x_{13} \,|\, \mathcal{M}_k)$.

Supplemental Exercise 14.1.2: Describe another one of the elementary points in the sample space for four future students.

Solution to Supplemental Exercise 14.1.2:

The sample space, according to Feller's definition, consists of $n^r = 16^4 = 65,536$ possibilities for four distinguishable balls labeled as **a, b, c, d** distributed over sixteen cells. For us, there are $M = 4$ students distinguishable by their names who possess a particular characterization on graduation status and three tests.

The most detailed description of an event in this sample space would be something like, ball **b** is in cell 2, balls **a** and **d** are in cell 13, and ball **c** is in cell 16. In our scenario, students Alex and Dawn did not graduate even though they scored high on tests C and D, Beth graduated with high scores on tests B and D, and, finally, poor Carl did not graduate while managing to score low on all three tests.

Supplemental Exercise 14.1.3: What is the probability of the event where two students share the same characterization, while the other two are different between themselves as well as the first two students?

Solution to Supplemental Exercise 14.1.3:

The previous exercise described just one of the elementary events comprising the compound event asked for in this exercise. Moving up the hierarchy, we know that there are twelve ways for the four students to distribute themselves over cells 2, 13, and 16. This answer we get from the multiplicity factor,

$$W(M) = \frac{M!}{M_1!\, M_2!\, \cdots\, M_{16}!} = \frac{4!}{0!\, 2!\, \cdots\, 1!\, 0!\, 0!\, 1!} = 12$$

Moving up to the next level, the contingency table,

0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1

illustrating the distribution of $r = 4$ balls over $n = 16$ cells is just one example of the $1,680$ possibilities for the integer partition of the sum $M = 4$,

$$M = 2 + 1 + 1 + \overbrace{0 + \cdots + 0}^{13} = 4$$

This number is found from,

$$\frac{n!}{r_z!\, r_s!\, \cdots\, r_M!} = \frac{16!}{13!\, 2!\, 1!\, 0!\, 0!} = 1680$$

Multiplying these results from the two counting formulas, we find that the total number of elementary points in the compound event is then $1680 \times 12 = 20,160$.

If "Maxwell–Boltzmann statistics" are used, that is, if the one model inserting the minimal amount of information into the probability distribution is assumed true, then $Q_i = 1/16$. The probability of the event is then calculated as,

$$\begin{aligned} P(\textbf{Compound Event}) &= \sum_{\text{Events}}^{1680} W(M)\, Q_1^{M_1}\, Q_2^{M_2}\, \cdots\, Q_{16}^{M_{16}} \\ &= 1680 \times [\ 12 \times (1/16)^0\, (1/16)^2\, \cdots\, (1/16)^1\ ] \\ &= 1680 \times \frac{12}{16^4} \\ &= \frac{20160}{65536} \end{aligned}$$

The probability for the event that all four students have a dissimilar characterization is the highest of all at,

$$\begin{aligned} P(\textbf{Compound Event}) &= \sum_{\text{Events}}^{1820} W(M)\, Q_1^{M_1}\, Q_2^{M_2}\, \cdots\, Q_{16}^{M_{16}} \\ &= 1820 \times [\ 24 \times (1/16)^0\, (1/16)^1\, \cdots\, (1/16)^1\ ] \\ &= 1820 \times \frac{24}{16^4} \\ &= \frac{43680}{65536} \end{aligned}$$

Thus, it is very likely for the compound event to occur where the all students have a different characterization, or at most two students share the same characterization,

$$P(\textbf{Compound Event}) = \frac{63840}{65536} = 0.9741$$

On the other hand, contrary to the adoption of a single model, if the IP cannot in good conscience claim that much knowledge about the model space, then it must resort to "Bose–Einstein statistics" where the prior probability over all conceivable models is the same. The consequence is that the probability for the above compound event is now lower at,

$$P(\textbf{Compound Event}) = \frac{3500}{3876} = 0.9030$$

It doesn't make any sense outside the realm of quantum physics, but if for some reason, two or more students were not permitted to share the same graduation test score characterization, then only the integer partition of M,

$$M = 1 + 1 + 1 + 1 + \overbrace{0 + \cdots + 0}^{12} = 4$$

would be allowed. Then, the so–called "Fermi–Dirac statistics" would allocate equal probability to each of the 364 possibilities where the four students share none of the characteristics,

$$P(\textbf{Event}) = \binom{n}{r}^{-1} = \left[\frac{n!}{r!\,(n-r)!}\right]^{-1} = \frac{4!\,12!}{16!} = \frac{1}{364}$$

14.2 Data Become Available

After some data are observed, the subtleties discussed in the previous Chapter begin to fade into the background. The degree of belief in future frequency counts will be driven more and more by the already observed frequency counts. Even so, please do remember that the formula used for these posterior predictive probabilities depends critically on assuming that the IP was in a completely uninformed state of knowledge about the relative standing of all the models existing in the model space before the arrival of the data. The data mentioned in these exercises are the same as shown in the contingency table of Figure 14.1, Volume I.

Supplemental Exercise 14.2.1: What is the probability of observing four future students graduate with low scores on all three tests given the data?

Solution to Supplemental Exercise 14.2.1:

The total number of future frequency counts is $\sum_{i=1}^{16} M_i = M = 4$ with $M_8 = 4$ and all other $M_i = 0$. Cell 8 of the joint probability table indexes the joint statement "A student GRADUATES with a LOW score on Test 1, a LOW score on Test 2, and a LOW score on Test 3." Applying the formula for the posterior predictive probability with $M = 4$, $N = 32$, and $n = 16$,

$$\begin{aligned}
P(M_1 = 0, \cdots, M_8 = 4, \cdots, M_{16} = 0 \,|\, \mathcal{D}) &= \frac{M!\,(N+n-1)!}{\prod_{i=1}^{16} N_i!\,(M+N+n-1)!} \times \frac{\prod_{i=1}^{16}(M_i+N_i)!}{\prod_{i=1}^{16} M_i!} \\
&= \frac{4!\,(32+16-1)!}{8!\cdots 5!\,(4+32+16-1)!} \times \frac{(0+8)!\cdots(4+2)!\cdots(0+5)!}{0!\cdots 4!\cdots 0!} \\
&= \frac{47!}{8!\cdots 5!\,51!} \times 8!\cdots 6!\cdots 5! \\
&= \frac{6\times 5\times 4\times 3}{51\times 50\times 49\times 48} \\
&= 6.0\times 10^{-5}
\end{aligned}$$

Intuitively, we might have expected such an event to have a low probability, and the above calculation confirms this feeling with a quantitative answer.

Supplemental Exercise 14.2.2: What is the probability of the data?

Solution to Supplemental Exercise 14.2.2:

It would be legitimate to raise the question: Why does one need to calculate the probability of something that is a given? In our current problem, we are always calculating the posterior predictive probability $P(M_1, M_2, \cdots, M_n \,|\, \mathcal{D})$. The data $\mathcal{D} \equiv N_2, N_2, \cdots, N_n$ have already happened. The easiest answer, but perhaps not the most satisfying, is simply that the formal manipulation rules require it.

In the most generic expression of Bayes's Theorem,

$$P(A \,|\, B) = \frac{P(A, B)}{P(B)}$$

B has happened as well, but we still have to find its probability. Having digested this, one could ask the follow–on question: Why is the **Product Rule** correct?

The probability of the data is a marginal sum over model space,

$$P(\mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(\mathcal{D} \,|\, \mathcal{M}_k)\; P(\mathcal{M}_k)$$

For a continuous model space, this expression gets translated into,

$$P(N_1, N_2, \cdots, N_n) = \int \cdot \int_{\sum_{i=1}^{n} q_i = 1} P(N_1, N_2, \ldots, N_n \,|\, q_1, q_2, \cdots, q_n)\; \text{pdf}\; (q_1, q_2, \cdots, q_n)\; \mathrm{d}q_i$$

The first term on the right hand side under the integral is a multinomial probability,

$$P(N_1, N_2, \ldots, N_n \,|\, q_1, q_2, \cdots, q_n) = W(N) \times q_1^{N_1}\; q_2^{N_2} \cdots q_n^{N_n}$$

The second term on the right hand side under the integral is the prior probability for a model. Operationally, this is implemented by a Dirichlet distribution with n α_i parameters.

$$\text{pdf}\; (q_1, q_2, \cdots, q_n) = C_D \times q_1^{\alpha_1 - 1}\; q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1}$$

Substituting for these two terms, we have,

$$P(N_1, N_2, \cdots, N_n) = \int \cdot \int_{\sum_{i=1}^{n} q_i = 1} W(N) \times q_1^{N_1}\; q_2^{N_2} \cdots q_n^{N_n} \times C_D \times q_1^{\alpha_1 - 1}\; q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1}\; \mathrm{d}q_i$$

Bring out from under the integral the two constant terms that do not depend on the q_i,

$$P(N_1, N_2, \cdots, N_n) = W(N) \times C_D \times \int \cdot \int_{\sum_{i=1}^n q_i = 1} q_1^{N_1}\, q_2^{N_2} \cdots q_n^{N_n} \times q_1^{\alpha_1 - 1}\, q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1}\, \mathrm{d}q_i$$

Add the exponents for each q_i term,

$$P(N_1, N_2, \cdots, N_n) = W(N) \times C_D \times \int \cdot \int_{\sum_{i=1}^n q_i = 1} q_1^{N_1 + \alpha_1 - 1}\, q_2^{N_2 + \alpha_2 - 1} \cdots q_n^{N_n + \alpha_n - 1}\, \mathrm{d}q_i$$

The known analytical solution to the multiple integral appears as the third term,

$$P(N_1, N_2, \cdots, N_n) = W(N) \times C_D \times \frac{\prod_{i=1}^n \Gamma(N_i + \alpha_i)}{\Gamma\left(\sum_{i=1}^n N_i + \alpha_i\right)}$$

This solution permits us to find the constant factor C_D for the Dirichlet distribution because of the universal constraint that all assigned probabilities must sum to 1,

$$\int \cdot \int_{\sum_{i=1}^n q_i = 1} C_D \times q_1^{\alpha_1 - 1}\, q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_n - 1}\, \mathrm{d}q_i = 1$$

Substituting this result brings us to,

$$P(N_1, N_2, \cdots, N_n) = W(N) \times \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^n \Gamma(N_i + \alpha_i)}{\Gamma\left(\sum_{i=1}^n N_i + \alpha_i\right)}$$

The right hand side becomes somewhat more transparent with,

$$\sum_{i=1}^n \alpha_i = \mathcal{A}, \quad \sum_{i=1}^n N_i = N$$

Also insert the factorial expression for the multiplicity factor $W(N)$ into the above expression,

$$P(N_1, N_2, \cdots, N_n) = \frac{N!}{N_1!\, N_2! \cdots N_n!} \times \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^n \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^n \Gamma(N_i + \alpha_i)}{\Gamma(N + \mathcal{A})}$$

It would be easy enough to code this last expression into *Mathematica* for any computation involving general α_i, but we will extend the derivation to cover the specific case of Laplace's **Principle of Insufficient Reason**. Here, all $\alpha_i = 1$, so we can greatly simplify the probability of the data to,

$$P(N_1, N_2, \cdots, N_n) = \frac{N!}{N_1!\, N_2!\, \cdots\, N_n!} \times (n-1)! \times \frac{N_1!\, N_2!\, \cdots\, N_n!}{(N+n-1)!}$$

$$= \frac{N!\,(n-1)!}{(N+n-1)!}$$

Both N and n are fixed for a given problem, so the probability for any possible pattern of the data is a constant. For example, the probability of the actual data in this example is,

$$P(N_1 = 8, N_2 = 2, \cdots, N_{16} = 5) = \frac{32!\,15!}{(32+16-1)!} = 1.33 \times 10^{-12}$$

However, this answer must be exactly the same as the probability for any other conceivable set of data like $P(N_1 = 32, \text{all remaining } N_i = 0)$. This is why we are so sanguine about writing the posterior probability as

$$P(\mathcal{M}_k \,|\, \mathcal{D}) \propto P(\mathcal{D} \,|\, \mathcal{M}_k) P(\mathcal{M}_k)$$

Supplemental Exercise 14.2.3: With the groundwork of the last exercise in hand, complete the derivation of the probability of future frequency counts given some past data.

Solution to Supplemental Exercise 14.2.3:

The general formula for the posterior predictive probability of future events given some already observed data for any state of knowledge about the causes of those events is derived in detail below. The major constructs involved in the proof have already been developed in the above derivation for the probability of the data.

The beginning step, as is so very often the case, is just to straightforwardly write out Bayes's Theorem for the problem as stated. Here, for the probability of future frequency counts given some past data, Bayes's Theorem would look like this,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) = \frac{P(M_1, M_2, \cdots, M_n, N_1, N_2, \cdots, N_n)}{P(N_1, N_2, \cdots, N_n)}$$

This is the same as Bernardo & Smith's formula previously quoted in section 12.2.

Written out in this manner, we can see immediately that we have already solved for the denominator,

$$P(\mathcal{D}) \equiv P(N_1, N_2, \cdots, N_n)$$

and all that's left is the numerator which follows the same steps very closely.

The key, and somewhat surprising ingredient, is the pleasant cancellation of terms in the numerator that takes place when the result for the probability of the data is inserted into the proof.

Focusing our attention, then, on the numerator of Bayes's Theorem, we have the familiar use of the **Sum Rule** to marginalize over all of the models,

$$P(M_1, M_2, \cdots, M_n, N_1, N_2, \cdots, N_n) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, \cdots, M_n, N_1, N_2, \cdots, N_n, \mathcal{M}_k)$$

Through the **Product Rule**, the right hand side under the summation becomes,

$$P(M_1, M_2, \cdots, M_n, N_1, N_2, \cdots, N_n, \mathcal{M}_k) = P(M_1, M_2, \cdots, M_n \,|\, \mathcal{M}_k) \times \\ P(N_1, N_2, \cdots, N_n \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k)$$

After substituting the multinomial formulas for the past and future frequency counts, the Dirichlet distribution for the prior probability of the models, and then embedding all of this within an integration where the region of integration must satisfy the constraint that $\mathcal{R} = \sum_{i=1}^{n} q_i = 1$,

$$P(M_1, M_2, \cdots, M_n, N_1, N_2, \cdots, N_n) = \int \cdot \int_{\mathcal{R}} W(M) \times q_1^{M_1} q_2^{M_2} \cdots q_n^{M_n} \times \\ W(N) \times q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n} \times \\ C_D \times q_1^{\alpha_1 - 1} q_2^{\alpha_2 - 1} \cdots q_n^{\alpha_1 - 1} \, \mathrm{d}q_i$$

As we have done so many times before, the next step involves pulling out all of the constants, and adding the exponents for the q_i,

$$P(M_1, \cdots, N_n) = W(M) \times W(N) \times C_D \times \\ \int \cdot \int_{\mathcal{R}} q_1^{M_1 + N_1 + \alpha_1 - 1} q_2^{M_2 + N_2 + \alpha_2 - 1} \cdots q_n^{M_n + N_n + \alpha_n - 1} \, \mathrm{d}q_i$$

Leave the first three terms as they are for the time being, and concentrate solely on the solution of the Dirichlet integral,

$$\int \cdot \int_{\mathcal{R}} q_1^{M_1 + N_1 + \alpha_1 - 1} q_2^{M_2 + N_2 + \alpha_2 - 1} \cdots q_n^{M_n + N_n + \alpha_n - 1} \, \mathrm{d}q_i = \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\Gamma\left(\sum_{i=1}^{n} M_i + N_i + \alpha_i\right)} \\ = \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\Gamma\left(M + N + \mathcal{A}\right)}$$

Now return to the C_D term,

$$C_D = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} = \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)}$$

We have set up the numerator for the cancellation that will occur when the denominator is brought back in,

$$P(M_1, \cdots, N_n) = W(M) \times W(N) \times \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\Gamma\,(M + N + \mathcal{A})}$$

We return to where we began at Bayes's Theorem,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) = \frac{P(M_1, M_2, \cdots, M_n, N_1, N_2, \cdots, N_n)}{P(N_1, N_2, \cdots, N_n)}$$

but now we can substitute in all of our new transformations. Reviewing the result for the denominator, the probability of the data was found as,

$$P(N_1, N_2, \cdots, N_n) = W(N) \times \frac{\Gamma\,(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{n} \Gamma(N_i + \alpha_i)}{\Gamma\,(N + \mathcal{A})}$$

Substituting our most recent results for the denominator yields,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) =$$

$$\frac{W(M) \times W(N) \times \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{n} \Gamma(M_i+N_i+\alpha_i)}{\Gamma\,(M+N+\mathcal{A})}}{W(N) \times \frac{\Gamma\,(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{n} \Gamma(N_i+\alpha_i)}{\Gamma\,(N+\mathcal{A})}}$$

Here is where we can take advantage of the aforementioned convenient cancellation of the second and third terms in the numerator by the corresponding terms in the denominator,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) = \frac{W(M) \times \frac{\prod_{i=1}^{n} \Gamma(M_i+N_i+\alpha_i)}{\Gamma\,(M+N+\mathcal{A})}}{\frac{\prod_{i=1}^{n} \Gamma(N_i+\alpha_i)}{\Gamma\,(N+\mathcal{A})}}$$

Or better yet,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) = W(M) \times \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\Gamma\,(M + N + \mathcal{A})} \times \frac{\Gamma\,(N + \mathcal{A})}{\prod_{i=1}^{n} \Gamma(N_i + \alpha_i)}$$

For the final step, collect all the terms which will be constants for any particular problem, in other words, $\sum_{i=1}^{n} M_i = M$, $\sum_{i=1}^{n} N_i = N$, $\sum_{i=1}^{n} \alpha_i = \mathcal{A}$,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n) = C \times \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\prod_{i=1}^{n} \Gamma(N_i + \alpha_i)}$$

where the constant multiplier C is,

$$C = W(M) \times \frac{\Gamma(N + \mathcal{A})}{\Gamma(M + N + \mathcal{A})}$$

The familiar steps in this derivation have been touched on previously at various levels of detail in Exercises 20.9.1, 22.6.8, 22.6.9, 32.9.8 in Volume II, and Exercises 47.8.5 and 47.8.6 in Volume III.

Supplemental Exercise 14.2.4: Write a *Mathematica* program to perform the calculations required for the posterior predictive probability.

Solution to Supplemental Exercise 14.2.4:

It is possible to write a very short program in *Mathematica* in order to carry out these calculations of the posterior predictive probability. Conveniently name this function **posteriorPredictiveProbability[** *arg1, arg2, arg3* **]** and provide it with three arguments.

The first argument is a list of the future frequency counts, the second argument is a list of the data, and the third and final argument is a list of the α_i parameters of the Dirichlet distribution.

```
posteriorPredictiveProbability[future_List, data_List,
                                alpha_List] :=
        Module[{largeN, M, A, term1, term2, term3},
            M = Total[future];
            largeN = Total[data];
            A = Total[alpha];
            term1 = Apply[Multinomial, future];
            term2 = Apply[Times, Gamma[future + data + alpha]];
            term3 = Gamma[largeN + alpha] /
                        Apply[Times, Gamma[data + alpha]];
            ScientificForm[N[term1 term2 term3], 7]]
```

Supplemental Exercise 14.2.5: Rely on the above program to check the hand calculation of the posterior predictive probability carried out in Supplemental Exercise 14.2.1.

Solution to Supplemental Exercise 14.2.5:

That exercise asked for the posterior predictive probability of seeing four future students with a particular set of traits. This is the list for the first argument. The data provided by 32 students were given in the contingency table shown in Figure 14.1 of Volume I. This is the list for the second argument. The α_i parameters reflect a uniform distribution. This is the list for the third argument.

The answer returned was 6.002401×10^{-5} after evaluating,

```
posteriorPredictiveProbability[
            {0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0},
            {8, 2, 3, 1, 1, 3, 4, 2, 0, 1, 0, 1, 0, 1, 0, 5},
            {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1}]
```

Supplemental Exercise 14.2.6: Rely on the above program to check the posterior predictive probability found at the end of section 14.2.2 in Volume I.

Solution to Supplemental Exercise 14.2.6:

There we asked the question of whether the IP should have a higher degree of belief in future frequency counts that mimicked where most of the data occurred. Cells 1, 6, 7, and 16 of the contingency table containing the data showed the largest past frequency counts.

Using **`posteriorPredictiveProbability[]`** and changing only the future frequency list in the first argument to

```
{1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1}
```

resulted in a probability of 4.321729×10^{-3}. (Exercise 14.5.5, Volume I).

This probability is over two times larger than, say, the situation where all four future students end up in cell 1, (Exercise 14.5.4, Volume I). This cell indexes the joint statement where a student graduates with HIGH scores on all three tests. This was the most frequent outcome as revealed by the data with eight students in cell 1.

Supplemental Exercise 14.2.7: Finish Geisser's example that was started in Supplemental Exercise 12.3.6.

Solution to Supplemental Exercise 14.2.7:

Even though Geisser's Example 3.5 was only supposed to illustrate de Finetti's theorem, he went on to find the formula for the posterior predictive probability. For us, this is exactly the topic of this Chapter, namely the formula,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots, N_n)$$

as applied to finding the probability for any number of future frequency counts of college students describing their test and graduation status given some past data.

Since de Finetti's theorem was explained in the context of a state space with dimension of $n = 2$, fall back onto the coin tossing scenario for a numerical example. We will use Geisser's formula to calculate the probability for seeing 5 future HEADS and 4 future TAILS in 9 future coin tosses given that the coin has already been tossed 100 times with 50 HEADS and 50 TAILS observed as the past data.

The probability computed for this event will then be compared with the prior predictive formula of the same event when no data were available and the IP was completely uninformed about the entire physical set of causes surrounding the coin

toss. This was in fact the answer of 1/10 found during the course of working out Supplemental Exercise 12.3.6.

Geisser derives a formula for the posterior predictive probability as shown below in his notation,

$$\Pr[R = r \,|\, t] = \int \binom{M}{r} \theta^r \, (1-\theta)^{M-r} \, p(\theta \,|\, x^{(N)}) \, d\theta$$

$$= \frac{\Gamma(M+1)\,\Gamma(N+\alpha+\beta)\,\Gamma(r+t+\alpha)\,\Gamma(M+N-r-t+\beta)}{\Gamma(r+1)\,\Gamma(M-r+1)\,\Gamma(\alpha+t)\,\Gamma(N-t+\beta)\,\Gamma(M+N+\alpha+\beta)}$$

It is easier to confirm Geisser's version by translating over into my notation given the extensive experience we have had with it. Review Supplemental Exercise 14.2.3 to refresh your memory. At the end of Supplemental Exercise 12.3.6, we had derived Geisser's final term $p(\theta \,|\, x^{(N)})$, the posterior probability for the models which for us is $P(\mathcal{M}_k \,|\, \mathcal{D})$.

Geisser defines $R = \sum_{i=1}^{M} X_{N+i}$, so it is clear that his M also denotes the total number of future frequency counts of successes and failures. The $X_1, X_2, \cdots, X_N$ represent the N available data in terms of past frequency counts. Thus, r is the number of future successes, $M - r$ is the number of future failures. Likewise, t is the number of past successes, and $N - t$ the number of past failures. The left hand side is then translated into,

$$\Pr[R = r \,|\, t] \equiv P(M_1, M_2 \,|\, N_1, N_2)$$

The first term under the integral is equivalent to,

$$\binom{M}{r} \theta^r \, (1-\theta)^{M-r} \equiv \frac{M!}{M_1!\, M_2!} \, q^{M_1} \, (1-q)^{M_2}$$

The entire expression on the right hand side is then,

$$\int \binom{M}{r} \theta^r \, (1-\theta)^{M-r} \, p(\theta \,|\, x^{(N)}) \, d\theta \equiv \sum_{k=1}^{\mathcal{M}} \frac{M!}{M_1!\, M_2!} \, q^{M_1} \, (1-q)^{M_2} \times P(\mathcal{M}_k \,|\, \mathcal{D})$$

$$\equiv \sum_{k=1}^{\mathcal{M}} \frac{M!}{M_1!\, M_2!} \, q^{M_1} \, (1-q)^{M_2} \times \frac{P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)}{P(\mathcal{D})}$$

$$\equiv \sum_{k=1}^{\mathcal{M}} \frac{M!}{M_1!\, M_2!} \, q^{M_1} \, (1-q)^{M_2} \times \frac{1}{P(\mathcal{D})} \times P(\mathcal{D} \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

Make the transition over to a continuous model space, and recall the important cancellation provided by $P(\mathcal{D})$,

$$
\begin{aligned}
P(M_1, M_2 \,|\, N_1, N_2) &= \int_0^1 \frac{M!}{M_1!\, M_2!}\, q^{M_1}\, (1-q)^{M_2} \times \frac{1}{P(\mathcal{D})} \times \frac{N!}{N_1!\, N_2!} q^{N_1} (1-q)^{N_2} \times \text{pdf}\,(q)\, \mathrm{d}q \\
&= \frac{M!}{M_1!\, M_2!} \times \frac{\Gamma(M_1+N_1+\alpha)\,\Gamma(M_2+N_2+\beta)}{\Gamma(M+N+\alpha+\beta)} \times \frac{\Gamma(N+\alpha+\beta)}{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)} \\
&= \frac{\Gamma(M+1)}{\Gamma(M_1+1)\,\Gamma(M-M_1+1)} \times \frac{\Gamma(M_1+N_1+\alpha)\,\Gamma(M_2+N_2+\beta)}{\Gamma(M+N+\alpha+\beta)} \times \\
&\quad \frac{\Gamma(N+\alpha+\beta)}{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)}
\end{aligned}
$$

It is possible to match up these terms in my derivation with Geisser's terms. Examine how Geisser took the final step from the posterior probability expression derived in Exercise 12.3.6 to the posterior predictive formula shown above. At the end of this exercise, we had reproduced Geisser's derivation of the posterior probability over model space in his notation as,

$$
p(\theta \,|\, x^{(N)}) = \frac{\Gamma(N+\alpha+\beta)\, \theta^{t+\alpha+1}\, (1-\theta)^{N-t+\beta-1}}{\Gamma(t+\alpha)\,\Gamma(N-t+\beta)}
$$

Following the template of,

$$
P(M_1, M_2 \,|\, \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k \,|\, \mathcal{D})
$$

Geisser's notation is translated into,

$$
\begin{aligned}
P(M_1, M_2 \,|\, \mathcal{M}_k) &= \frac{M!}{M_1!\, M_2!}\, q^{M_1}\, (1-q)^{M_2} \\
&\equiv \binom{M}{r} \theta^r (1-\theta)^{M-r} \\
P(\mathcal{M}_k \,|\, \mathcal{D}) &\equiv p(\theta \,|\, x^{(N)}) \\
P(M_1, M_2 \,|\, \mathcal{D}) &\equiv \int \binom{M}{r} \theta^r (1-\theta)^{M-r} \times \frac{\Gamma(N+\alpha+\beta)\, \theta^{t+\alpha+1}\, (1-\theta)^{N-t+\beta-1}}{\Gamma(t+\alpha)\,\Gamma(N-t+\beta)}\, d\theta
\end{aligned}
$$

We now commence the final series of transformations,

$$Pr[R = r \,|\, t] = \int \binom{M}{r} \theta^r (1-\theta)^{M-r} \times$$

$$\frac{\Gamma(N+\alpha+\beta)\,\theta^{t+\alpha+1}\,(1-\theta)^{N-t+\beta-1}}{\Gamma(t+\alpha)\,\Gamma(N-t+\beta)}\,d\theta$$

$$= \binom{M}{r} \frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)} \times$$

$$\int \theta^r (1-\theta)^{M-r} \times \theta^{t+\alpha+1}\,(1-\theta)^{N-t+\beta-1}\,d\theta$$

$$\binom{M}{r} = \frac{\Gamma(M+1)}{\Gamma(r+1)\,\Gamma(M-r+1)}$$

$$\int \theta^r (1-\theta)^{M-r} \times \theta^{t+\alpha+1}\,(1-\theta)^{N-t+\beta-1}\,d\theta = \int \theta^{r+t+\alpha-1}(1-\theta)^{M-r+N-t+\beta-1}\,d\theta$$

$$= \frac{\Gamma(r+t+\alpha)\,\Gamma(M+N-r-t+\beta)}{\Gamma(M+N+\alpha+\beta)}$$

Put all the pieces back together,

$$\Pr[R = r \,|\, t] = \frac{\Gamma(M+1)\,\Gamma(N+\alpha+\beta)\,\Gamma(r+t+\alpha)\,\Gamma(M+N-r-t+\beta)}{\Gamma(r+1)\,\Gamma(M-r+1)\,\Gamma(\alpha+t)\,\Gamma(N-t+\beta)\,\Gamma(M+N+\alpha+\beta)}$$

Now go back to my derivation of the posterior predictive which ended up with,

$$P(M_1, M_2 \,|\, N_1, N_2) = \frac{\Gamma(M+1)}{\Gamma(M_1+1)\,\Gamma(M-M_1+1)} \times \frac{\Gamma(M_1+N_1+\alpha)\,\Gamma(M_2+N_2+\beta)}{\Gamma(M+N+\alpha+\beta)}$$

$$\times \frac{\Gamma(N+\alpha+\beta)}{\Gamma(N_1+\alpha)\,\Gamma(N_2+\beta)}$$

Since,

$$M_1 \equiv r$$

$$M_2 \equiv M - r$$

$$N_1 \equiv t$$

$$N_2 \equiv N - t$$

$$P(M_1, M_2 \,|\, N_1, N_2) = \frac{\Gamma(M+1)\,\Gamma(N+\alpha+\beta)\,\Gamma(r+t+\alpha)\,\Gamma(M-r+N-t+\beta)}{\Gamma(r+1)\,\Gamma(M-r+1)\,\Gamma(\alpha+t)\,\Gamma(N-t+\beta)\,\Gamma(M+N+\alpha+\beta)}$$

so that in the final analysis

$$P(M_1, M_2 \,|\, N_1, N_2) = Pr[R = r \,|\, t]$$

The main goal of this exercise was to ensure that I carried out an independent confirmation of my posterior predictive probability formula. Geisser's derivation in his Example 3.5 provides such an independent confirmation. I guess my not very rigorous mathematical rationale is simply that it is highly unlikely that two different long series of transformations would end up with the same expression unless they were not saying the same thing.

Supplemental Exercise 14.2.8: Let *Mathematica* compute the posterior predictive probability for obtaining five HEADS and four TAILS in nine future coin tosses using my formula and Geisser's formula.

Solution to Supplemental Exercise 14.2.8:

This was actually the goal of the last exercise because we wanted to see how the formal rules updated a degree of belief in some future event starting from Laplace's uninformed state as compared to an informed state based solely on past data. Since my formula for the posterior predictive probability has already been coded in,

```
posteriorPredictiveProbability[
                              future_List, data_List, alpha_List]
```

evaluating

`posteriorPredictiveProbability[{5, 4}, {50, 50}, {1, 1}]` returns the posterior predictive probability,

$$P(M_1 = 5, M_2 = 4 \,|\, N_1 = 50, N_2 = 50) = 0.236934$$

Code Geisser's formula directly into *Mathematica* with,

```
Geisser[M_, largeN_, r_, t_, alpha_, beta_] :=
   N[(Gamma[M + 1] * Gamma[largeN + alpha + beta] *
      Gamma[r + t + alpha] * Gamma[M + largeN - r - t + beta]) /
      (Gamma[r + 1] * Gamma[M - r +1] * Gamma[alpha + t] *
      Gamma[largeN - t + beta] * Gamma[M + largeN + alpha + beta])]
```

Evaluating `Geisser[9, 100, 5, 50, 1, 1]` also returns $\Pr[\,R = r \,|\, t\,] = 0.236934$. We thus have numerical confirmation that the two posterior predictive formulas are computing the same probabilities.

Of far more interest to me, however, is to quantitatively assess the "deleterious" impact of an uninformed prior probability over model space when compared to an informed prior probability. As many Bayesian commentators have noted over the years, with increasing data, the impact of Laplace's uninformed prior probability recedes into insignificance.

As a numerical example, suppose an IP possessed prior knowledge about the coin such that it was willing to change the α and β parameters of the *beta distribution* from Laplace's value of 1, to, say, a value of 3. This change indicates a willingness to

believe more in a fair coin. The *beta distribution* as a prior probability is unimodally centered around an assignment of $q = 1/2$ tapering off to higher and lower values for q. Laplace's uninformed prior probability does not favor any q over the entire interval from 0 to 1.

Nonetheless, running,

```
posteriorPredictiveProbability[{5, 4}, {50, 50}, {3, 3}]
```

only raises the posterior predictive probability for obtaining five HEADS to 0.237263. It does remain true that the more informative prior probability will achieve the same probability as the uninformed prior probability with fewer data points.

14.3 Ratio of Posterior Probabilities for Models

Supplemental Exercise 14.3.1: Write a short *Mathematica* program to calculate the ratio of posterior probabilities based on the development of section 14.3.1 of Volume I.

Solution to Supplemental Exercise 14.3.1:

Create a user–defined function **ratioOfPosteriorProbabilities[***arg1, arg2***]** with two arguments. The first argument is a list of the data. The second argument is a list of the probability assignments under Model B.

```
ratioOfPosteriorProbabilities[data_List, qB_List] :=
              Module[{qA, largeN, freq, entropy },
                       qA = Table[1 / 16, {1, 16}];
                       largeN = Total[data];
                       freq = data / largeN;
                       entropy = Total[freq * Log[qA / qB]];
                       Exp[largeN * entropy]]
```

Supplemental Exercise 14.3.2: Find the ratio of posterior probabilities where Model A is the fair model and Model B hews very closely to the actual data.

Solution to Supplemental Exercise 14.3.2:

This was the example examined in Exercise 14.5.8 of Volume I. The first argument is a list of the data given in the contingency table shown in Figure 14.1 of Volume I. The second argument is whatever model probabilities, as encapsulated in Model B, we might be interested in.

Here, they are the same as the normed frequency counts except for Q_{1B} which

has 0.0004 subtracted from the normed frequency count in order to make up for the assignments of 0.0001 to Q_{9B}, Q_{11B}, Q_{13B}, and Q_{15B}. We wanted to avoid a division by zero error message in the entropy expression. The probability assignments **qA** under Model A were hardwired in as the familiar assignments under the fair model.

Evaluating,

```
ratioOfPosteriorProbabilities[
        {8, 2, 3, 1, 1, 3, 4, 2, 0, 1, 0, 1, 0, 1, 0, 5},
        {8/32 -.0004, 2/32, 3/32, 1/32, 1/32, 3/32, 4/32, 2/32,
          .0001, 1/32, .0001, 1/32, .0001, 1/32, .0001, 5/32}]
```

returns the value of 2.77886×10^{-8} calculated in Volume I. This ratio of posterior model probabilities reveals the severely reduced influence of the fair model in any predictions about future events. Prior to the data, the fair model had just as much influence on predicting the probability for the first occurrence of any statement as any other model.

In Exercise 14.5.9 of Volume I, another model, $\mathcal{M}_C$, whose assignments were closer to the data than the fair model of $\mathcal{M}_A$, was proposed. Make a minor change to **ratioOfPosteriorProbabilities[ ]** by explicitly including **qC** as a third argument to the function, and adjust the rest of the program as necessary.

With a third argument the list of probabilities under model $\mathcal{M}_C$,

```
ratioOfPosteriorProbabilities[
        {8, 2, 3, 1, 1, 3, 4, 2, 0, 1, 0, 1, 0, 1, 0, 5},
        {8/32 -.0004, 2/32, 3/32, 1/32, 1/32, 3/32, 4/32, 2/32,
          .0001, 1/32, .0001, 1/32, .0001, 1/32, .0001, 5/32},
        {8/32 -.04, 2/32, 3/32, 1/32, 1/32, 3/32, 4/32, 2/32,
          .01, 1/32, .01, 1/32, .01, 1/32, .01, 5/32}]
```

returns the value of 3.98293 calculated in Volume I. Model B's assignments still make a greater contribution to the average than Model C's, but with nowhere near the relative strength as compared to Model A.

14.4 Averaging Over All Models

These examples in Volume I averaged over just two models. What happens if the average is taken over *all* models? The probability that the very next student has any particular pattern of traits will always be modulated downwards when the average is taken with respect to all models.

For example, the last exercise for Chapter 14, Volume I, Exercise 14.5.10, asked for the impact on the ratio of posterior probabilities for models B and C if the data had been slightly different. Model B's relative strength over Model C was reduced to 2.819 for the particular small change in the data examined there.

The conditional probability for GRADUATES when given the test scores on all three tests for the next student was reduced to 0.7408 when averaging over just these two models. However, if this same conditional probability is calculated by averaging with respect to all models, then the probability is modulated downwards to 0.70.

Supplemental Exercise 14.4.1: What do Bayes's Theorem together with the posterior predictive probability tell us about the probability that the very next student will graduate given high scores on all three tests?

Solution to Supplemental Exercise 14.4.1:

This exercise is similar to Ex 14.5.10 of Volume I in that the data in cells 1 and 9 have been changed to $N_1 = 6$ and $N_9 = 2$. Write out Bayes's Theorem for the next student to graduate with joint probabilities conditioned on the data appearing on the right hand side,

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{D}) =$$

$$\frac{P(A_{N+1}, B_{N+1}, C_{N+1}, D_{N+1}, \,|\, \mathcal{D})}{P(A_{N+1}, B_{N+1}, C_{N+1}, D_{N+1}, \,|\, \mathcal{D}) + P(\overline{A}_{N+1}, B_{N+1}, C_{N+1}, D_{N+1}, \,|\, \mathcal{D})}$$

But consider: these joint probabilities for the next student are exactly what we find with the posterior predictive probability. For the numerator and first term in the denominator, $M = 1$, $M_1 = 1$ and all the remaining $M_i = 0$. For the second term in the denominator, $M = 1$, $M_9 = 1$ and all the remaining $M_i = 0$. The posterior predictive probabilities for the case where $M = 1$ reduces to,

$$P(M_1 = 1, \cdots, M_9 = 0, \cdots, M_{16} = 0 \,|\, \mathcal{D}) = \frac{N_1 + 1}{48}$$

$$P(M_1 = 0, \cdots, M_9 = 1, \cdots, M_{16} = 0 \,|\, \mathcal{D}) = \frac{N_9 + 1}{48}$$

Substituting back into Bayes's Theorem with the changed data of $N_1 = 6$ and $N_9 = 2$,

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, D_{N+1}, \mathcal{D}) = \frac{(6+1)}{(6+1) + (2+1)} = 0.70$$

Supplemental Exercise 14.4.2: Revisit the general symbolic formula for the posterior predictive probability presented in Supplemental Exercise 14.2.3 to verify the numbers in the last exercise.

Solution to Supplemental Exercise 14.4.2:

We will verify the probability that appeared in the numerator of Baye's Theorem in the last exercise. With $M = 1$, $M_1 = 1$, $N = 32$, $n = 16$, and $\mathcal{A} = 16$, we have,

$$P(M_1, \cdots, M_{16} \,|\, \mathcal{D}) = W(M) \times \frac{\Gamma(N+\mathcal{A})}{\Gamma(M+N+\mathcal{A})} \times \frac{\prod_{i=1}^{16} \Gamma(M_i + N_i + \alpha_i)}{\prod_{i=1}^{16} \Gamma(N_i + \alpha_i)}$$

$$W(M) = 1$$

$$\Gamma(N+\mathcal{A}) = 47!$$

$$\Gamma(M+N+\mathcal{A}) = 48! = (M+N+n-1)!$$

$$\frac{\Gamma(N+\mathcal{A})}{\Gamma(M+N+\mathcal{A})} = \frac{1}{48}$$

$$\prod_{i=1}^{16} \Gamma(M_i + N_i + \alpha_i) = (N_1+1)!\; N_2!\; \cdots\; N_{16}!$$

$$\prod_{i=1}^{16} \Gamma(N_i + \alpha_i) = N_1!\; N_2!\; \cdots\; N_{16}!$$

$$\frac{\prod_{i=1}^{16} \Gamma(M_i + N_i + \alpha_i)}{\prod_{i=1}^{16} \Gamma(N_i + \alpha_i)} = N_1 + 1$$

$$P(M_1 = 1, \cdots, M_{16} = 0 \,|\, \mathcal{D}) = \frac{N_1 + 1}{M + N + n - 1} = \frac{7}{48}$$

This answer can be confirmed by evaluating,

```
posteriorPredictiveProbability[
            {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {6, 2, 3, 1, 1, 3, 4, 2, 2, 1, 0, 1, 0, 1, 0, 5},
            {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1}]
```

which returns $1.458333 \times 10^{-1} = 7/48$.

Chapter 15

What Does Uninformed Mean?

15.1 Extending the Kangaroo Scenario

The basic kangaroo inferential scenario made its initial appearance in Volume I, Chapter Fifteen. With a state space of dimension $n = 4$, this scenario served to illustrate the effect on the probabilities of future frequency counts when the Dirichlet distribution parameters, the four α_i, were systematically altered.

Let's extend the numerical investigation of the impact of these α_i parameters on our understanding of what it means for an IP to be "uninformed." To that end, the basic kangaroo scenario is bumped up by considering three statements involving a kangaroo's physical and psychological traits. Add a kangaroo's fur color to its beer and hand preference in order to enlarge the dimension of the state space to $n = 8$. Furthermore, enlarge the number of future frequency counts from $M = 16$ kangaroos of the basic scenario to $M = 80$ kangaroos. This enhanced inferential scenario is quite familiar from its constant use in both Volumes II and III.

Any contingency table or joint probability table will now consist of eight cells. One special contingency table we will examine is where all 80 kangaroos are evenly spread out over the eight cells, ten to each cell. Another is where all 80 kangaroos are crammed into just one cell, say, cell 6. Just as in Table 15.1 of Volume I, the probability of these future frequency counts will be calculated as all eight α_i parameters are increased in lockstep starting at all $\alpha_i = 1$.

Of course, the main conceptual breakthrough we are trying to achieve is a better understanding of the role that information plays when it is inserted into the prior probability over model space, $P(\mathcal{M}_k)$. This information is directly implemented through the α_i parameters of the Dirichlet distribution. When all $\alpha_i = 1$, the IP is implementing Laplace's description of an uninformed IP's epistemological state as it bears on the causes of events.

Supplemental Exercise 15.1.1: Calculate the probabilities of the future frequency counts for these two special contingency tables as mentioned in the Introduction when the parameters in the Dirichlet distribution are altered.

Solution to Supplemental Exercise 15.1.1:

Table 15.1 below is constructed to correspond to Table 15.1 in Volume I. The first column now shows the common value of the parameter for all eight α_i rather than listing all eight individually. The notation $\mathcal{A}$ is the sum $\sum_{i=1}^{8} \alpha_i$. The third column is the probability when there are 10 kangaroos in each cell of the contingency table. The next column is the probability when all 80 kangaroos are in cell 6, and no kangaroos are in any of the other seven cells of the contingency table. And, as before, the rounded ratio of these two probabilities is shown in the final column.

Table 15.1: *The probability of future frequency counts for two special eight cell contingency tables when the α_i parameters for the prior probability of a model are changed.*

α_i	$\mathcal{A}$	$P(\text{all } M_i = 10)$	$P(M_6 = 80)$	*Ratio*
1	8	1.71×10^{-10}	1.71×10^{-10}	1
2	16	1.94×10^{-9}	7.34×10^{-16}	2.65×10^{6}
5	40	2.64×10^{-8}	5.05×10^{-26}	5.22×10^{17}
10	80	1.15×10^{-7}	1.38×10^{-35}	8.34×10^{27}
20	160	3.22×10^{-7}	1.33×10^{-45}	2.42×10^{38}
35	280	5.57×10^{-7}	4.66×10^{-53}	1.19×10^{46}
∞	∞	1.35×10^{-6}	5.66×10^{-73}	2.38×10^{66}

When the IP is completely uninformed about the causes of events, in other words, when all the $\alpha_i = 1$, the first row shows that the probability of any possible pattern of future frequency counts is going to be the same.

As the α_i values slowly increase, the special contingency table with ten counts in all eight cells becomes more and more probable. The special contingency table with all eighty counts in any one cell becomes vastly more improbable. The ratio of these two probabilities in the final column illustrates how the information for the prior probability of models affects the IP's degree of belief in obtaining these contingency tables.

The final row shows how the asymptotic value for the probability of the first special set of future frequency counts has increased, and the probability for the second special set of future frequency counts has decreased.

Supplemental Exercise 15.1.2: Calculate the probability for ten counts in each cell of the contingency table when only one model has become dominant. Use the multinomial probability distribution.

Solution to Supplemental Exercise 15.1.2:

When all eight parameters of the Dirichlet distribution for the probability of the models approach infinity, then only one model is operative. This model is,

$$\mathcal{M}_k \to (q_1 = 1/8, q_2 = 1/8, \cdots, q_8 = 1/8)$$

Applying the multinomial probability distribution to the first special contingency table where all eight cells contain ten kangaroos,

$$\begin{aligned} P(M_1 = 10, M_2 = 10, \cdots, M_8 = 10) &= W(M) \times Q_1^{M_1} \times Q_2^{M_2} \times \cdots \times Q_8^{M_8} \\ &= \frac{80!}{10!\,10!\,\cdots\,10!} \times (1/8)^{10} \times (1/8)^{10} \times \cdots (1/8)^{10} \\ &= (2.38 \times 10^{66}) \times (5.66 \times 10^{-73}) \\ &= 1.35 \times 10^{-6} \end{aligned}$$

the same finding as in the final row of the table in the first supplemental exercise. Notice the notational change from q_i to Q_i indicating that the integration over model space has taken place.

Supplemental Exercise 15.1.3: Once again, just as in the above exercise, use the multinomial probability distribution to calculate the probability for eighty counts in cell 6 of the contingency table when only one model has become dominant.

Solution to Supplemental Exercise 15.1.3:

When all eight parameters of the Dirichlet distribution for the probability of the models approach infinity, then only one model is operative. Obviously, this model must be the same as in the previous exercise,

$$\mathcal{M}_k \to (q_1 = 1/8, q_2 = 1/8, \cdots, q_8 = 1/8)$$

Applying the multinomial probability distribution for the second special contingency table where all eighty kangaroos are contained in cell 6,

$$
\begin{aligned}
P(M_1 = 0, \cdots, M_6 = 80, \cdots, M_8 = 0) &= W(M) \times Q_1^{M_1} \times Q_2^{M_2} \times \cdots \times Q_8^{M_8} \\
&= \frac{80!}{0! \cdots 80! \cdots 0!} \times (1/8)^0 \cdots \times (1/8)^{80} \cdots \times (1/8)^0 \\
&= 1 \times (5.66 \times 10^{-73}) \\
&= 5.66 \times 10^{-73}
\end{aligned}
$$

again, the same result as in the final row of the table in the first supplemental exercise, 15.1.1.

Supplemental Exercise 15.1.4: Offer an intuitive interpretation of the impact of the multiplicity factor from these numerical results.

Solution to Supplemental Exercise 15.1.4:

The vast change evident in the probabilities of the two special contingency tables is encapsulated in the multiplicity factor $W(M)$. The ratio of the probabilities for the two contingency tables approaches $W(M)$ as the α_i parameters approach ∞.

Compare this to the situation when the IP is completely uninformed about the causes for events. Then, the multiplicity factor for different contingency tables has no effect whatsoever. $W(M)$ for ten counts in all eight cells remains the huge factor 2.38×10^{66} and $W(M)$ for all eighty counts in cell 6 remains at 1. But when all $\alpha_i = 1$, this huge disparity in the values of the multiplicity factor doesn't count; the probability for each contingency table is equal.

As the $\alpha_i \to \infty$, the IP abandons Laplace's epistemological stance of total ignorance about the causes of events. It becomes more and more certain that it does know the cause for events, and in this particular case, that known cause is reflected in the model,

$$\mathcal{M}_k \to (q_1 = 1/8, q_2 = 1/8, \cdots, q_8 = 1/8)$$

Then the multiplicity factor takes over and provides the quantitative disparity in the degree of belief of seeing all eighty kangaroos evenly distributed over the set of traits versus all eighty possessing the same trait.

Supplemental Exercise 15.1.5: What *Mathematica* code computed the values in Table 15.1?

Solution to Supplemental Exercise 15.1.5:

The following very direct *Mathematica* program calculated the values appearing in Table 15.1 of Supplemental Exercise 15.1.1.

```
priorPredictiveSETable151[alpha_] := Module[
              {future1, future2, M, alphalist, A, Cterm, e1, e2},
     future1 = {10, 10, 10, 10, 10, 10, 10, 10};
     future2 = {0, 0, 0, 0, 0, 80, 0, 0};
     M = Total[future1];
     alphalist = Table[alpha, {i, 1, 8}];
     A = Total[alphalist];
     Cterm = M! Gamma[A] / Gamma[M + A];
     e1 = N[Cterm Apply[Times, (Gamma[future1 + alphalist]) /
               (future1! Gamma[alphalist])]];
     e2 = N[Cterm Apply[Times, (Gamma[future2 + alphalist]) /
               (future2! Gamma[alphalist])]];
     {ScientificForm[e1, 3], ScientificForm[e2, 3],
      ScientificForm[e1 / e2, 3]}]
```

After this workhorse code becomes available, it is easy to have *Mathematica* generate a table,

```
TableForm[Table[priorPredictiveSETable151[alpha],
                                {alpha, {1, 2, 5, 10, 20, 35}}]]
```

illustrating how the entries to Table 15.1 were filled in. The final row where all $\alpha_i \to \infty$ was computed by the multiplicity factor,

$$W(M) = \frac{80!}{10!\ 10!\ \cdots\ 10!} = 2.38 \times 10^{66}$$

Supplemental Exercise 15.1.6: How can you make *Mathematica* compute the probability in Exercise 15.1.2?

Solution to Supplemental Exercise 15.1.6:

I made use of the **Multinomial[ ]** function to compute the multiplicity factor and the **Product[ ]** function to compute the $Q_i^{M_i}$ terms.

```
Multinomial[10, 10, 10, 10, 10, 10, 10, 10] *
                  Product[Power[1/8, 10], {i, 1, 8}] // N
```

Supplemental Exercise 15.1.7: How many possible contingency tables are there in total for this enhanced kangaroo scenario?

Solution to Supplemental Exercise 15.1.7:

The formula for computing the number of contingency tables was given in Equation (15.1) of Volume I as,

$$\textbf{total contingency tables} = \frac{(M+n-1)!}{M!\,(n-1)!}$$

which works out to nearly six billion possibilities for $M = 80$ and $n = 8$,

$$\begin{aligned}\textbf{total contingency tables} &= \frac{87!}{80!\,7!} \\ &= 5,843,355,957\end{aligned}$$

How many contingency tables from this above total of nearly six billion have all 80 kangaroos crammed into a single cell? There are obviously eight such possibilities. How many contingency tables have 79 kangaroos crammed into one cell and a single kangaroo in one of the remaining cells? It might not be so obvious that there are now 56 such contingency tables. One of these 56 possibilities is that 79 kangaroos are sandy fur colored, right handed, Corona drinkers. The remaining kangaroo is a beige fur colored, left handed, Foster's drinker.

Supplemental Exercise 15.1.8: What is the formula for calculating the number of contingency tables in the previous exercise?

Solution to Supplemental Exercise 15.1.8:

In Exercise 15.7.2 of Volume I we used this formula for finding out the number of various contingency tables when a die was rolled four times with $M = 4$ and $n = 6$,

$$\textbf{contingency tables} = \frac{n!}{r_z!\,r_s!\,r_d!\cdots r_M!}$$

Applying this same formula with $M = 80$ and $n = 8$, we find that for 80 kangaroos in one of the eight cells, and no kangaroos in the remaining seven cells, there is a repetition count $r_{80} = 1$, a repetition count of $r_z = 7$, with all of the remaining repetition counts equal to zero. Thus,

$$\textbf{contingency tables} = \frac{8!}{7!\,0!\cdots 0!\,1!} = 8$$

For 79 kangaroos in one of the eight cells, and one kangaroo in one of the remaining seven cells there is a repetition count $r_{80} = 0$, $r_{79} = 1$, $r_1 = 1$, and, finally, a repetition count of $r_z = 6$, with all of the remaining repetition counts equal to zero. Thus,

$$\textbf{contingency tables} = \frac{8!}{6!\,1!\cdots 1!\,0!} = 56$$

How many contingency tables would show that the kangaroos are evenly spread out over the three traits? There can be only one such contingency table with ten kangaroos found in each of the eight cells of the contingency table,

$$\textbf{contingency tables} = \frac{8!}{0!\,0!\cdots r_{10} = 8!\cdots 0!} = 1$$

Then where are all of the other contingency tables since there are nearly six billion of them? Any pattern showing just one repetition of a frequency count where all eight frequency counts add up to eighty will result in $8! = 40,320$ contingency tables. For example, suppose there are frequency counts of 1 through 7 somewhere in the contingency table with a frequency count of 52 to make up the total of $M = 80$ frequency counts in whatever cell remains. There are 40,320 such tables,

$$\textbf{contingency tables} = \frac{8!}{0!\,1!\,1!\cdots r_{52} = 1!\cdots 0!} = 40,320$$

For another example, suppose there are frequency counts of 2 through 8 somewhere in the contingency table with a frequency count of 45 to make up the total of $M = 80$ frequency counts in whatever cell remains. There are again 40,320 such tables,

$$\textbf{contingency tables} = \frac{8!}{0!\,0!\,1!\,1!\cdots r_{45} = 1!\cdots 0!} = 40,320$$

Supplemental Exercise 15.1.9: How many contingency tables are there when the only constraint is that 73 kangaroos are in one cell?

Solution to Supplemental Exercise 15.1.9:

Table 15.2 at the top of the next page lists fifteen different contingency tables satisfying the constraint that 73 kangaroos are in one cell. But each one of these contingency tables can be formed in a number of ways according to the formula we have been using. For example, one integer partition of 80, shown in row 13 as $73 + 2 + 2 + 1 + 1 + 1 + 0 + 0$ can occur in 1,680 contingency tables,

$$\textbf{contingency tables} = \frac{8!}{2!\,3!\,2!\,0!\cdots r_{73} = 1!\cdots 0!} = 1,680$$

So now it becomes easier to conceive of how nearly six billion contingency tables might arise. *Mathematica* will provide you with the listing of all fifteen integer partitions through,

```
Column[IntegerPartitions[80, {8}, {0, 1, 2, 3, 4, 5, 6, 7, 73}]]
```

Table 15.2: *A listing of 13,728 contingency tables from the total of nearly six billion that contain a frequency count of 73 kangaroos in one cell.*

Row	*Integer partition*	*Number*
1	$73+7+0+0+0+0+0+0$	56
2	$73+6+1+0+0+0+0+0$	336
3	$73+5+2+0+0+0+0+0$	336
4	$73+5+1+1+0+0+0+0$	840
5	$73+4+3+0+0+0+0+0$	336
6	$73+4+2+1+0+0+0+0$	1680
7	$73+4+1+1+1+0+0+0$	1120
8	$73+3+3+1+0+0+0+0$	840
9	$73+3+2+2+0+0+0+0$	840
10	$73+3+2+1+1+0+0+0$	3360
11	$73+3+1+1+1+1+0+0$	840
12	$73+2+2+2+1+0+0+0$	1120
13	$73+2+2+1+1+1+0+0$	1680
14	$73+2+1+1+1+1+1+0$	336
15	$73+1+1+1+1+1+1+1$	8
Total		13728

Supplemental Exercise 15.1.10: How many elementary points are there in the sample space for this inferential scenario?

Solution to Supplemental Exercise 15.1.10:

If the numbers above weren't large enough, then the number of elementary points in the sample space staggers the imagination. There are a total of $n^M = 8^{80} \approx 1.77 \times 10^{72}$ elementary points for this rather modest scenario. Nonetheless, it is relatively easy to drill down and imagine what pattern any elementary point must assume.

From the total of 5,843,355,957 contingency tables, and since we already have the numbers from the previous exercise, imagine it is one of the 56 contingency tables with 73 kangaroos in one cell with the remaining 7 kangaroos in another cell. There are no kangaroos in the other six cells. This situation is represented by the first row of Table 15.2. Refer back to Figure 22.1 in Volume II for a sketch of the eight cell joint probability table to index the cell number to the trait.

Now select one of these 56 contingency tables meeting this criterion. Suppose we select the contingency table where 73 kangaroos are in cell 3, 7 kangaroos are in cell 6, with no kangaroos in cells 1, 2, 4, 5, 7, and 8. In other words, 73 kangaroos were observed to be sandy right handed Corona drinkers, 7 kangaroos to be beige left handed Foster's drinkers, and no kangaroos were observed to possess any of the other six traits.

But we are not yet down to the level of an elementary point. Each of the 80 kangaroos is an individual kangaroo with a name. We would have to provide the specific names of the 7 beige left handed Foster's drinkers and the 73 sandy right handed Corona drinkers to finally specify one elementary point in the sample space consisting of a total of 1.77×10^{72} elementary points.

The multiplicity factor tells us how many possibilities there are for a frequency counts of 7 in cell 6 and 73 in cell 3. There are

$$W(M) = \frac{M!}{M_1!\, M_2! \cdots M_8!} = \frac{80!}{0!\, 0!\, 73!\, 0!\, 0!\, 7!\, 0!\, 0!} = 3,176,716,400$$

a little over three billion elementary points even after we have drilled this far down to a specific contingency table.

As a very rough idea of the order of magnitude contributed by contingency tables with 73 kangaroos in one cell to the overall number of elementary points, we have $10^4 \times 10^9 = 10^{13}$, somewhere in the general vicinity of tens of trillions.

Supplemental Exercise 15.1.11: What point am I trying to make with all of these numerical exercises?

Solution to Supplemental Exercise 15.1.11:

What is the moral I wish to convey with all of these rather boring numerical exercises involving fantastically large numbers?

It seems to me that Feller's choice of elementary points in a sample space as the fundamental foundation for conceptualizing probability is ill–advised. I say "ill–advised" because it runs up against these fantastically large numbers very, very quickly. Our kangaroo scenario is by all accounts quite modest in scope with a state space constructed from only three binary statements accompanied by an equally modest sample size.

Thinking about probabilities should not become mired in confusion in thinking about fantastically large numbers. Can anyone really justify a probability as an IP's quantitative degree of belief in the individual truth of any one of these 10^{72} elementary points? Well, that's what an IP is doing when it claims from Feller's sampling space perspective that, for example,

$$P(M_1 = 10, M_2 = 10, \cdots, M_8 = 10) = \frac{W(M)}{n^M} = \frac{\frac{80!}{10!\,10! \cdots 10!}}{8^{80}} = 1.35 \times 10^{-6}$$

All $8^{80} \approx 1.77 \times 10^{72}$ elementary points have the same probability, and there are $W(M)$ elementary points that satisfy the condition of ten kangaroos to each cell of the contingency table.

I would like you to pay attention to the conceptual distinction in calculating this probability from the sample space perspective as was done above when compared

to the way I think the probability should be conceptualized. There are only eight probabilities, q_1 through q_8, the numerical assignments to each joint statement in the state space. Certainly this is a small enough number that doesn't immediately make our heads spin.

If the prior probability over model space gets constructed because the IP inserted information reflected by all the $\alpha_i \to \infty$, then each $q_i = 1/8$. As a consequence, the numerical assignment to the probability for each of the eight joint statements about a kangaroo's traits is $Q_i = 1/8$.

The formal manipulation rules for probability expressions then dictate, starting out for example with the first two kangaroos as right handed Corona drinkers, with the first beige fur colored, and the second sandy fur colored,

$$\begin{aligned} P(A_2 = a_5, A_1 = a_7, \mathcal{M}_k) &= P(A_2 \,|\, A_1, \mathcal{M}_k) \times P(A_1 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \\ &= P(A_2 = a_5 \,|\, \mathcal{M}_k) \times P(A_1 = a_7 \,|\, \mathcal{M}_k) \times P(\mathcal{M}_k) \\ &= q_5 \times q_7 \times P(\mathcal{M}_k) \end{aligned}$$

Eventually, the probability expression for all eighty kangaroos gets built up into,

$$P(A_{80} = a_5, A_{79} = a_7, \cdots, A_1 = a_4, \mathcal{M}_k) = \overbrace{q_5 \times q_7 \times q_2 \times \cdots q_4}^{\text{80 terms}} \times P(\mathcal{M}_k)$$

Through the **Commutativity** property of multiplication, after rearranging and collecting all eighty terms this becomes,

$$P(A_{80} = a_5, A_{79} = a_7, \cdots, A_1 = a_4, \mathcal{M}_k) = q_1^{M_1} \times q_2^{M_2} \times \cdots \times q_8^{M_8} \times P(\mathcal{M}_k)$$

If the specific names of the kangaroos involved in the data are not important, then the multiplicity factor can be used to construct the final probability,

$$P(M_1 = 10, M_2 = 10, \cdots, M_8 = 10) = W(M) \times (1/8)^{10} \times (1/8)^{10} \times \cdots \times (1/8)^{10}$$

In arriving at this expression, the prior probability for the models had to be inserted, and then integrated over all possible values of q_i.

We achieve the same answer as Feller's sample space approach. However, any extreme numbers that might appear on the right hand side arise not from a sample space, but rather from applying the rules of probability. Therefore, the origin of these numbers, however extreme they may become, is not in any way mysterious.

But discussing the trade–off in order to arrive at the same answer is critical. First of all, the impact of the models had to be marginalized out of the expression on the left hand side,

$$\begin{aligned} P(M_1, M_2, \cdots, M_n) &= W(M) \int \cdot \int \prod_{i=1}^{n} q_i^{M_i} \text{ pdf } (q_i)\, dq_i \\ &= W(M) \int \cdot \int \prod_{i=1}^{n} q_i^{M_i}\, \delta(q_i - 1/n)\, dq_i \\ &= W(M) \times \prod_{i=1}^{n} (1/n)^{M_i} \\ &= \frac{W(M)}{n^M} \end{aligned}$$

When the Dirichlet distribution is used for the prior probability, pdf(q_i) and the parameters of the distribution are set so that all $\alpha_i \to \infty$, the density function for the prior probability turns into a Dirac delta function which, in turn, simplifies the integral into the multinomial distribution.

This information inserted into the prior probability is in some sense the polar opposite of what Laplace inserts under "complete ignorance." The IP is insisting that one model, and one model only, may assign numerical values to probabilities for the joint statements. And that one model assigns the specific numerical value of $1/n$ to all of the joint statements in the state space. Laplace, by letting all the α_i parameters equal 1, asserts instead that all models are treated on an equal basis. Every assignment from 1s and 0s to $1/n$, and everything in between must be averaged through the integration.

15.2 Uninformed About What?

Chapter Fifteen of Volume I took on the task of examining what it meant for an IP to be "uninformed." It turns out, as Jaynes masterfully uncovered after pondering this perennial epithet, that there can be more than one way to characterize what it means for an IP to be completely uninformed. The previous section explored one way that the IP can be uninformed.

And this first characterization focused on an IP's maximized uncertainty about the model space. An IP implements an operational definition for being completely uninformed about the relative standing of all the models by setting all of the α_i parameters of the Dirichlet probability density function equal to 1.

As a consequence, the prior probability for every model $P(\mathcal{M}_k)$ is one and the same, or, in other words, the prior probability density function is "flat." This tactic will serve to fulfill the original Bayes–Laplace explanation leading to Bayes's Theorem.

But, through what Jaynes labeled as an "uncertainty relationship," the fact that the IP is uninformed about the relative standing for all of the models DOES NOT imply that the IP is therefore completely uninformed about the elementary points in the sample space. This fact is clearly evident with our choice of the special contingency tables.

The multiplicity factor is nullified when the IP is completely uninformed about the models. All 80 kangaroos with the same trait is represented by only eight elementary points in the sample space, while there are about 2.38×10^{66} elementary points in the sample space when the kangaroos are evenly spread out over all eight traits. And yet, the probability for these two polar opposite future frequency counts is exactly the same under an IP's total state of ignorance. As a consequence, the IP cannot be "uninformed" about the relative importance of these elementary points.

We illustrated Jaynes's beautiful relationship by allowing the α_i parameters to incrementally march in lockstep away from their initial value of 1 towards a value of infinity. Now, the IP becomes more and more certain about the models until, at the culmination, it is willing to place all of its bets on just one model that assigns a numerical value of $1/n$ to each joint statement.

The multiplicity factor is restored to its former glory. The contingency table with 10 kangaroos to each cell of the contingency table is $W(M)$ times more likely than having all 80 kangaroos crammed into one cell. We are back to Feller's initial description of probabilities assigned to elementary points in the sample space where each of the n^M elementary points has the same weight.

Having absorbed these concepts, and relying once again on the structure of the Dirichlet distribution and its parameters for additional insight, Jaynes was led to inquire about the mirror analogy of letting all the α_i parameters start off at 1, but now march off towards to 0. Numerical exploration of these cases leads one to conceive of another, arguably even stranger, way to characterize an IP as "uninformed."

Supplemental Exercise 15.2.1: Construct the analog to Table 15.2 of Volume I for the enhanced kangaroo scenario.

Solution to Supplemental Exercise 15.2.1:

The first column of Table 15.3 shown at the top of the next page lists five special contingency tables. The first row is where all 80 kangaroos are evenly spread out over all eight traits. The next four rows show all 80 kangaroos concentrated in cells 1, 4, 6, and 7 respectively. The next column starts off the march towards 0 with all eight $\alpha_i = 0.5$.

Recall that all five of these special contingency tables have the same probability of 1.71×10^{-10} when they start off at $\alpha_i = 1$. Each succeeding column reduces the α_i until, at the final column, all $\alpha_i = 0.0001$.

Table 15.3: *The probability for five special contingency tables when all eight of the parameters for the prior probability approach 0.*

All eight $\alpha_i \to 0$					
Frequencies	0.5000	0.1000	0.0100	0.0010	0.0001
All $M_i = 10$	1.01×10^{-11}	2.54×10^{-15}	8.47×10^{-22}	$\approx 10^{-28}$	$\approx 10^{-35}$
$M_1 = 80$	6.85×10^{-7}	5.70×10^{-3}	8.88×10^{-2}	0.1207	0.1246
$M_4 = 80$	6.85×10^{-7}	5.70×10^{-3}	8.88×10^{-2}	0.1207	0.1246
$M_6 = 80$	6.85×10^{-7}	5.70×10^{-3}	8.88×10^{-2}	0.1207	0.1246
$M_7 = 80$	6.85×10^{-7}	5.70×10^{-3}	8.88×10^{-2}	0.1207	0.1246

These numerical results reveal the same pattern found before in Volume I. The probability for the contingency table that possesses the maximum multiplicity factor is steadily declining to an extremely low probability as the $\alpha_i \to 0$. Meanwhile, any contingency table that possesses the minimum multiplicity factor of 1 is approaching a value of $1/n = 1/8 = 0.1250$ as the $\alpha_i \to 0$.

This kind of behavior is exactly what I meant when I mentioned a "second, stranger version of an IP being uninformed." The IP is definitely committed to the idea of strong causality of the type that Jeffreys wanted when he saw the Sun rise every morning without fail, or hydrogen and oxygen always combining to form water. If something happens once, then under strong causality it should happen again and again without exception. If one kangaroo is a left handed, beige, fur colored Corona drinker, then all kangaroos should exhibit these same traits under the information in these models where the $\alpha_i \to 0$.

The uncertainty, or the precise nature of being uninformed about a kangaroo's traits, resides in the IP not knowing which strong causality model holds true. All of the eight contingency tables with 80 kangaroos crammed into one cell end up splitting the available probability evenly among themselves. Are all of the kangaroos right handed sandy fur colored Foster's drinkers, or are they all of one the other seven types? There is no probability left over for any other contingency table, and especially not for the contingency table with the maximum multiplicity factor.

15.3 Pólya's Urn Scheme

In Chapter Fifteen we surveyed some of the amazing consequences ensuing from adoption of the formal manipulation rule template that finds the probability for a statement by averaging over the prior probability for all of the models making

numerical assignments to the statement,

$$P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)$$

If the choice is made for technical convenience to capture the information in the above prior probability through the Dirichlet distribution with its alpha parameters, then many mysteries are cleared up. This was our primary goal in Chapter Fifteen, especially where it concerned that slippery concept of an IP who is "uninformed."

But what is even more amazing, and almost beyond belief, is that this same tactic works for the inferential scenario that has been labeled as *Pólya's Urn Scheme* in the probability theory literature. My motivation for including it here stems from the fact that the solution depends on fooling around with the alpha parameters of the Dirichlet distribution. In other words, the solution to this inference involves a slight variation of what was done in Chapter Fifteen.

My curiosity was initially piqued by the conceptual issues raised by reading Feller and trying to understand his presentation of de Finetti's Representation Theorem. You may recall that I have already broached this topic in the most cursory of ways in supplemental exercises under Chapter Twelve.

It is possible to paraphrase my canonical template as was given above into Feller's notation appropriate for de Finetti's Theorem as,

$$\left\{ P(A = a_1) = \sum_{k=1}^{\mathcal{M}} P(A = a_1 \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k) \right\} \equiv \left\{ P(X_1 = 1) = \int_0^1 \theta\, F\{\mathrm{d}\theta\} \right\}$$

After a long and obscure derivation of de Finetti's Theorem, Feller chooses to enlighten us with three examples supposedly illustrating an application of de Finetti's Theorem. He himself admits that his first example leads to "surprising results." And, boy, does it ever!

Here is Feller's presentation of the *Pólya urn model* (Volume II, pg. 229–230, Eq. (4.7)) illustrating, as said, an application of de Finetti's Theorem.

> **Examples.** (a) *In Pólya's urn model of* **1**; V, 2 an urn contains originally b black balls and r red balls. After each drawing the ball is returned and c balls of the color drawn are added to the urn. Thus the probability of a black ball in each of the first n drawings equals
>
> $$c_n = \frac{b\,(b+c)\cdots(b+(n-1)c)}{(b+r)\cdots(b+r+(n-1)c)} = \frac{\Gamma(\frac{b}{c}+n)\,\Gamma(\frac{b+r}{c})}{\Gamma(\frac{b+r}{c}+n)\,\Gamma(\frac{b}{c})}$$
>
> Put $\mathbf{X}_n = 1$ or 0 according as the nth drawing results in black or red. The easy calculation in **1**; V, 2 shows that these variables are exchangeable and hence c_n represents the nth moment of a distribution F. The appearance of [the above equation] reminds one of the beta integral II (2.5), and inspection shows that F *is the beta distribution II (4.2) with parameters* $\mu = b/c$ and $\nu =$

> r/c. Again using the beta integral it is seen that (4.1) [de Finetti's Theorem] agrees with **1**; V, (2.3) and (4.2) [de Finetti's Theorem with multiplicity factor] with **1**; V, (2.4). [Emphasis in the original.]

OK, so how can we deconstruct Feller's assertions as detailed in the above quote? Well, my preferred approach is to generate a numerical example and see if I am able to follow his recipe and come out with the correct answer. First, I have to find out if I can shoe horn Feller's notation into my own inferential notation where I do understand what is going on.

Supplemental Exercise 15.3.1: What is the probability of drawing three black balls from an urn consisting of seven black balls and three red balls?

Solution to Supplemental Exercise 15.3.1:

We are going to adopt Pólya's urn model just as Feller described it. Thus, we have $b = 7$ black balls, $r = 3$ red balls, with $c = 1$ ball of the same color as the ball drawn added to the urn. We are interested in the probability of drawing three black balls from this urn, so $M = 3$ with $M_1 = 3$ and $M_2 = 0$. Feller uses the notation of n for our M. Since there are only two possible observations at each draw, the state space is of dimension $n = 2$ in our notation for n.

Our standard prior predictive formula instructs us to solve this problem by first writing out,

$$P(M_1 = 3, M_2 = 0) = \int_0^1 W(M)\, q^{M_1}(1-q)^{M_2}\, \mathrm{pdf}\,(q)\, \mathrm{d}q$$

Immediately upon writing out this expression some doubt begins to creep in. Surely this template will not work for the inferential scenario where the probabilities are changing with each draw of a black or red ball from the urn?

Let's cast these doubts aside and plunge ahead to see where it leads us. The multiplicity factor $W(M) = 1$ disappears after the first line,

$$P(M_1 = 3, M_2 = 0) = \int_0^1 q^{M_1}(1-q)^{M_2}\, \mathrm{pdf}\,(q)\, \mathrm{d}q$$

Plug in the values for M_1 and M_2,

$$\begin{aligned} P(M_1 = 3, M_2 = 0) &= \int_0^1 q^3(1-q)^0\, \mathrm{pdf}\,(q)\, \mathrm{d}q \\ &= \int_0^1 q^3\, \mathrm{pdf}\,(q)\, \mathrm{d}q \end{aligned}$$

There is absolutely no issue with following Feller's advice that his F distribution is a *beta distribution*. We would do exactly the same. But critical is noticing that

we are leaving the information resident in the prior probability open for the time being, by not specifying values for α and β,

$$P(M_1 = 3, M_2 = 0) = \int_0^1 q^3 \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q$$

Continuing on with our standard operations, bring out C_{Beta} and add exponents,

$$P(M_1 = 3, M_2 = 0) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^1 q^{3+\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q$$

These operations are very familiar from our derivations of the prior and posterior predictive formulas.

Now, we arrive at the really interesting and surprising steps. Refer back to Feller's example to get an idea of what he is suggesting for the α and β parameters of the Dirichlet distribution. For α, he suggests $\mu = b/c = 7/1 = 7$ and for β, $\nu = r/c = 3/1 = 3$. Remember that Feller's $n = 3$ future drawings is our $M = 3$.

If we now go ahead and plug in these values for α and β into our last step, we have,

$$\begin{aligned} P(M_1 = 3, M_2 = 0) &= \frac{\Gamma(10)}{\Gamma(7)\,\Gamma(3)} \int_0^1 q^{3+7-1}\,(1-q)^{3-1}\,\mathrm{d}q \\ &= \frac{\Gamma(10)}{\Gamma(7)\,\Gamma(3)} \int_0^1 q^9\,(1-q)^2\,\mathrm{d}q \end{aligned}$$

Finish up with the solution to the integral,

$$\begin{aligned} P(M_1 = 3, M_2 = 0) &= \frac{\Gamma(10)}{\Gamma(7)\,\Gamma(3)} \int_0^1 q^9\,(1-q)^2\,\mathrm{d}q \\ &= \frac{\Gamma(10)}{\Gamma(7)\,\Gamma(3)} \times \frac{\Gamma(10)\times\Gamma(3)}{\Gamma(13)} \\ &= \frac{\Gamma(10)\,\Gamma(10)}{\Gamma(7)\,\Gamma(13)} \end{aligned}$$

So, at the end of all this legerdemain with the α and β parameters of the *beta distribution* as a prior probability over model space, we have a value for the supposed probability of drawing three black balls from the urn. And this answer was found by a straightforward application of our standard template for the probability of any future frequency counts. But how does it compare to Feller's answer?

Supplemental Exercise 15.3.2: What is Feller's solution?

Solution to Supplemental Exercise 15.3.2:

After plugging in the relevant values of $n = 3$, $b = 7$, $r = 3$, $c = 1$, and $\mu = b/c$, $\nu = r/c$ into Feller's Equation (4.7),

$$\begin{aligned} c_n &= \frac{\Gamma(\frac{b}{c} + n)\,\Gamma(\frac{b+r}{c})}{\Gamma(\frac{b+r}{c} + n)\,\Gamma(\frac{b}{c})} \\ &= \frac{\Gamma(\frac{7}{1} + 3)\,\Gamma(\frac{7+3}{1})}{\Gamma(\frac{7+3}{1} + 3)\,\Gamma(\frac{7}{1})} \\ &= \frac{\Gamma(10)\,\Gamma(10)}{\Gamma(13)\,\Gamma(7)} \end{aligned}$$

This is patently the same answer as we found in the previous exercise. The notation of c_n on the left hand side for which we wrote $P(M_1 = 3, M_2 = 0)$ stems from Feller's development of de Finetti's Theorem,

$$\begin{aligned} c_3 &= \int_0^1 \theta^3\, F\{\mathrm{d}\theta\} \\ &= \int_0^1 q^3\, \mathrm{pdf}\,(q)\,\mathrm{d}q \end{aligned}$$

It is the third moment of θ with respect to whatever distribution is represented by $F\{\mathrm{d}\theta\}$.

Recall that Feller is presenting this exercise as an application of de Finetti's Theorem. Does this mean that our standard template from the formal manipulation rules results in the same outcome as de Finetti's Theorem? It is hard to avoid this conclusion. But we will leave it for now as a conjecture that must be subjected to much critical appraisal.

Supplemental Exercise 15.3.3: What is the solution by referring back to basic principles?

Solution to Supplemental Exercise 15.3.3:

If we proceed from the direct statement of the problem as adhering to the dictates of the *Pólya Urn Scheme*, the probability for three black balls on the first three draws is,

$$
\begin{aligned}
P(B_3, B_2, B_1) &= P(B_3 \,|\, B_2, B_1) \times P(B_2 \,|\, B_1) \times P(B_1) \\
&= \frac{b+2}{b+r+2} \times \frac{b+1}{b+r+1} \times \frac{b}{b+r} \\
&= \frac{9}{12} \times \frac{8}{11} \times \frac{7}{10} \\
&= \frac{9!\,9!}{12!\,6!} \\
&= \frac{\Gamma(10)\,\Gamma(10)}{\Gamma(13)\,\Gamma(7)}
\end{aligned}
$$

the same as Feller's solution in supplemental exercise 15.3.2, and my solution in supplemental exercise 15.3.1.

It is seen that this direct writing out of the probabilities is the same as Feller's first expression for $c_n \equiv c_3$ as,

$$
c_n = \frac{b\,(b+c)\cdots(b+(n-1)\,c)}{(b+r)\cdots(b+r+(n-1)\,c}
$$

Supplemental Exercise 15.3.4: Offer up some reflections after digesting these results.

Solution to Supplemental Exercise 15.3.4:

Just as Feller warned us, these are truly surprising results. We seem to have reached an amazing conclusion that, even though the probability of a black ball is changing at every draw, we can still insert a constant probability θ at each draw as long as $F\{\mathrm{d}\theta\}$ is the *beta distribution* with parameters $\mu = 7$ and $\nu = 3$. De Finetti's Theorem goes through without a hitch.

Or, in my notation, the prior predictive probability of some number of future frequency counts can be correctly computed from the template suggested by the formal manipulation rule,

$$
P(A = a_i) = \sum_{k=1}^{\mathcal{M}} P(A = a_i \,|\, \mathcal{M}_k)\, P(\mathcal{M}_k)
$$

The analog to de Finetti's Theorem is simply,

$$
P(M_1, M_2) = \int_0^1 q^3 \times C_{\mathrm{Beta}}\, q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q
$$

The only mystery is why these particular values of $\alpha = 7$ and $\beta = 3$ are required in the prior probability over model space to arrive at the correct answer.

One interpretation might be to look at the shape of pdf (q). It is unimodal with its single peak at a mean at 0.7 but shading off gradually to 1 on the right side and to 0 on the left side. It is neither a flat distribution like the uninformative Laplace prior with $\alpha = \beta = 1$ nor a sharply peaked distribution captured by the Dirac δ–function, say, $(\delta - 0.7)$ when both α and $\beta \to \infty$ explored previously. Values of q around, say, between 0.5 and 0.8 will be preferentially chosen over values for q between 0 and 0.4 when the averaging over models takes place.

Apparently, the information inserted into the prior probability over model space reflected by these particular parameter values for the *beta distribution* is sufficient to offset the constant value of q at each draw, but it does seem so very contrived.

Supplemental Exercise 15.3.5: Rely upon *Mathematica* to plot the beta distribution with parameters $\alpha = 7$ and $\beta = 3$.

Solution to Supplemental Exercise 15.3.5:

Figure 15.1 shown below is what the preferred *beta distribution* used in de Finetti's Theorem for Feller's *Pólya's Urn Scheme* example looks like,

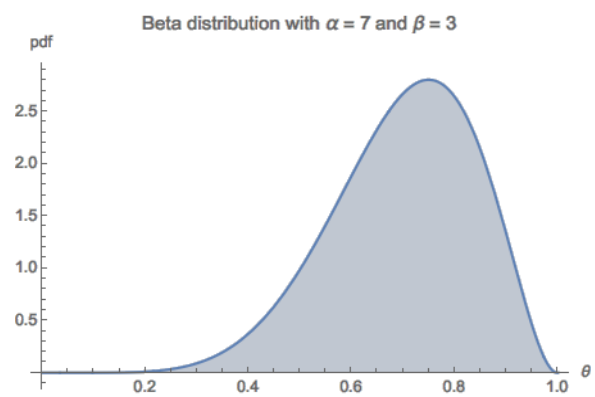

Figure 15.1: *A* Mathematica *plot of a beta distribution with parameters* $\alpha = 7$ *and* $\beta = 3$.

Supplemental Exercise 15.3.6: What is a simple Monte Carlo approach using *Mathematica* that approximates the exact computation of the probability for three black balls in the Pólya Urn Scheme?

Solution to Supplemental Exercise 15.3.6:

In effect, what we are asking for is an alternate verification of de Finetti's Theorem through a Monte Carlo calculation of the integral appearing in the theorem. The

Monte Carlo estimate of an integral is generically,

$$E[\,f(\theta)\,] = \int_{\mathcal{R}} f(\theta)\,\mathrm{pdf}\,(\theta)\,\mathrm{d}\theta \approx \sum_{i=1}^{N} f_i(\theta)/N$$

A Monte Carlo version of de Finetti's Theorem would involve a large random sample from a beta distribution with appropriate α and β parameters. Each jth sample would provide one $\theta_{(j)}$ and $(1-\theta)_{(j)}$. Sum $[\,\theta_{(j)}^{k} \times (1-\theta)_{(j)}^{n-k}\,]$ where both n and k are fixed for a particular problem over N such samples. Then divide by N to obtain a sample mean to estimate the integral in de Finetti's Theorem,

$$\frac{\sum_{j=1}^{N}[\,\theta_{(j)}^{k} \times (1-\theta)_{(j)}^{n-k}\,]}{N} \approx \int_0^1 \theta^k\,(1-\theta)^{n-k}\,F\{\mathrm{d}\theta\}$$

Since the $\theta_{(j)}$ have been randomly sampled from the appropriate *beta distribution*, $F\{\mathrm{d}\theta\}$, the proper weighting has been applied to each component in the summation to approximate the integral.

A direct instantiation into *Mathematica* of the Monte Carlo approximation to the probability of drawing three black balls is,

```
polya[largeN_] := Total[Power[RandomVariate[
                    BetaDistribution[7, 3], largeN]], 3] / largeN
```

Evaluating the de Finetti integral with 100,000 samples through **`polya[100 000]`** yields the approximation,

$$P(X_1 = 1, X_2 = 1, X_3 = 1) \approx 0.381503$$

The exact answer to the integral as calculated in the previous exercises is,

```
N[(Gamma[10] Gamma[10]) / (Gamma[7] Gamma[13])]
```

$$P(X_1 = 1, X_2 = 1, X_3 = 1) = 0.381818$$

Supplemental Exercise 15.3.7: Provide a detailed breakdown of what is happening in the above Monte Carlo approximation.

Solution to Supplemental Exercise 15.3.7:

A random selection of a legitimate value for θ is taken from the *beta distribution* with parameters $\alpha = 7$ and $\beta = 3$. Let's suppose that the first value produced by **`RandomVariate[BetaDistribution[7, 3]]`** is $\theta_{(1)} = 0.7682$.

This is then an assigned probability to drawing a black ball. The probability for three black balls is, according to the first term in de Finetti's Theorem,

$$\theta_{(1)} \times \theta_{(1)} \times \theta_{(1)} = 0.7682 \times 0.7682 \times 0.7682 = 0.4533$$

This value is then the starting value in a sum over very many repetitions of the above procedure.

When **`RandomVariate[BetaDistribution[7, 3]]`** is called for the second time, and say it produces a $\theta_{(2)} = 0.5617$, the value of,

$$\theta_{(2)} \times \theta_{(2)} \times \theta_{(2)} = 0.5617 \times 0.5617 \times 0.5617 = 0.1772$$

is added to 0.4533. Take a glance back at Figure 15.1, a plot of the probability density function of the *beta distribution*, to notice that both of these first two values are not surprising. Also, as a rough check, we know that $\theta_{(j)}$ values will tend to be randomly sampled around 0.7. $0.7^3 = 0.34$ is close to the correct probability of 0.3818.

After the last pick of $\theta_{(100,000)}$ from the beta distribution, the final sum was 38150.3. After dividing this sum by $N = 100,000$, the approximation to the actual value of the de Finetti integral of 0.381818 was found as 0.381503.

Perform the analytical integration with *Mathematica* returning the value of 0.381818 through,

```
N[Integrate[θ^3 PDF[BetaDistribution[7, 3], θ], {θ, 0, 1}]]
```

Supplemental Exercise 15.3.8: What data would have resulted in the same answer?

Solution to Supplemental Exercise 15.3.8:

Is the information in the prior probability over model space the same *as if* some data of previous drawings were available and the IP were uninformed before the data? If $N_1 = 7$ and $N_2 = 3$, is this the same as knowing the constituents of the urn, that is, there were exactly seven black balls and three red balls in the urn?

Thus, if the IP could have looked at each of the ten balls in the urn, with all of the balls then being replaced into the urn, do we end up with the same answer? Does knowledge of the exact contents of the urn solve the mystery?

Or, so I reasoned. But perhaps I was wrong about the assumption of the IP being "uninformed." When I actually started to do the computations this is what I found.

First, let me recapitulate the formula for calculating the probability of future drawings without conditioning on any N_i. This is the no data solution we found in the above exercises. Refer back to Exercise 15.7.9 in Volume I in order to locate an intermediate step in the derivation,

$$P(M_1, M_2, \cdots, M_n) = W(M) \times \frac{\Gamma(\mathcal{A})}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^{n} \Gamma(M_i + \alpha_i)}{\Gamma(M + \mathcal{A})}$$

With $n = 2$, $M = 3$, $M_1 = 3$, $M_2 = 0$, $\alpha_1 = \alpha = 7$, $\alpha_2 = \beta = 3$, $\mathcal{A} = 10$, the probability of drawing three black balls according to the *Pólya Urn Scheme* is,

$$\begin{aligned} P(M_1 = 3, M_2 = 0) &= W(3) \times \frac{\Gamma(10)}{\Gamma(7)\,\Gamma(3)} \times \frac{\Gamma(3+7)\,\Gamma(0+3)}{\Gamma(10+3)} \\ &= 1 \times \frac{\Gamma(10)}{\Gamma(7)\,\Gamma(3)} \times \frac{\Gamma(10)\,\Gamma(3)}{\Gamma(13)} \\ &= \frac{\Gamma(10)}{\Gamma(7)} \times \frac{\Gamma(10)}{\Gamma(13)} \\ &= \frac{9!\,9!}{6!\,12!} \end{aligned}$$

Now, transition to thinking about the problem as conditioned on some data N_i. After examining every single ball, noting the color, and finding the totals, replace all of the balls back into the urn. What would the data have to look like in order to reproduce the same answer as just calculated for the probability of drawing three black balls?

Look to the formula carefully derived in Supplemental Exercise 14.2.3,

$$P(M_1, M_2, \cdots, M_n \,|\, N_1, N_2, \cdots N_n) = C \times \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\prod_{i=1}^{n} \Gamma(N_i + \alpha_i)}$$

with the constant term,

$$C = W(M) \times \frac{\Gamma(N + \mathcal{A})}{\Gamma(M + N + \mathcal{A})}$$

With $n = 2$, $M = 3$, $M_1 = 3$, $M_2 = 0$, $\alpha_1 = \alpha = 1$, $\alpha_2 = \beta = 1$, $\mathcal{A} = 2$, $N = 10$, $N_1 = 7$, $N_2 = 3$, the probability of drawing three black balls according to the *Pólya Urn Scheme* changes to,

$$\begin{aligned} P(M_1 = 3, M_2 = 0 \,|\, N_1 = 7, N_2 = 3) &= C \times \frac{\Gamma(3+7+1) \times \Gamma(0+3+1)}{\Gamma(7+1) \times \Gamma(3+1)} \\ &= C \times \frac{\Gamma(11)}{\Gamma(8)} \\ C &= 1 \times \frac{\Gamma(10+2)}{\Gamma(3+10+2)} \\ P(M_1 = 3, M_2 = 0 \,|\, N_1 = 7, N_2 = 3) &= \frac{\Gamma(12)}{\Gamma(15)} \times \frac{\Gamma(11)}{\Gamma(8)} \end{aligned}$$

This is not the answer we were expecting. To gain some insight, change this answer into factorials to see what probabilities of a black ball are being assigned,

$$P(M_1 = 3, M_2 = 0 \,|\, N_1 = 7, N_2 = 3) = \frac{8}{12} \times \frac{9}{13} \times \frac{10}{14}$$

When compared to the no data solution,

$$P(M_1 = 3, M_2 = 0) = \frac{7}{10} \times \frac{8}{11} \times \frac{9}{12}$$

we see that we are off by 1 in the numerator and 2 in the denominator.

What change could we make to our input to correct this? We can't change n, M or the M_i. We could change the data to $N = 8$ with $N_1 = 6$ and $N_2 = 2$. This would bring us back to the correct answer. But this is kind of difficult to justify. Why would we examine only $N = 8$ of the balls in the urn instead of all $N = 10$ balls? Why would we leave exactly one black ball and one red ball in the urn?

There is one out left for us to explore. We could reduce the parameters of the *beta distribution* from $\alpha = \beta = 1$ to $\alpha = \beta = 0$. Numerically, everything works out correctly. We have managed to correct for the extra 1 in the numerator and the extra 2 in the denominator in the data based probability.

Apparently, we were not completely uninformed about the contents of the urn after examining all of its contents, to put it mildly. We seem to be pushed in the opposite direction of causality. But this rationale is no more satisfactory to me than the previous one.

Supplemental Exercise 15.3.9: Rely on *Mathematica* to compute the probabilities for all of the possibilities in three drawings from the Pólya Urn.

Solution to Supplemental Exercise 15.3.9:

We have concentrated on the numerical solution to drawing three black balls in the *Pólya Urn Scheme* because that is the example that Feller used. But we had better also compute all of the other possibilities and at least do the validity check that the probabilities sum to 1.

Drawing three balls in succession from the urn results in four macro–statements, or future frequency counts. The first is the one we have been discussing where all three black balls are drawn. The other three possibilities are 1) two black and one red, 2) two red and one black, and 3) all three red. There is only one way for the first and last possibilities to occur, but three ways for the second and third. These different multiplicity factors are accounted for in our prior predictive formula.

Adapting the prior predictive formula presented in the previous exercise for the $n = 2$ case yields,

$$P(M_1, M_2) = \frac{M!}{M_1!\,M_2!} \times \frac{\Gamma(\mathcal{A})}{\Gamma(\alpha)\,\Gamma(\beta)} \times \frac{\Gamma(M_1+\alpha)\,\Gamma(M_2+\beta)}{\Gamma(M+\mathcal{A})}$$

Separating out a constant term and reducing to two terms yields,

$$P(M_1, M_2) = \frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})} \times \frac{\Gamma(M_1+\alpha)\,\Gamma(M_2+\beta)}{M_1!\,M_2!\,\Gamma(\alpha)\,\Gamma(\beta)}$$

Implementing this formula in *Mathematica* we have a familiar looking function with two arguments, the first list containing the desired future frequency counts, and the second list containing the desired parameters for the *beta distribution*,

```
priorPredictive[future_List, alpha_List] := Module[{M, A, Cterm},
                M = Total[future];
                A = Total[alpha];
                Cterm = M! Gamma[A] / Gamma[M + A];
                N[Cterm Apply[Times,
                (Gamma[future + alpha] / future! Gamma[alpha])]
```

To verify our running example of $M_1 = 3$, $M_2 = 0$, $\alpha = 7$, and $\beta = 3$, evaluate,

```
priorPredictive[{3, 0}, {7, 3}]
```

The probability of drawing three black balls is $P(M_1 = 3, M_2 = 0) = 0.3818$. To calculate the probabilities for all four possibilities, evaluate,

```
Table[priorPredictive[{3 - i, i}, {7, 3}], {i, 0, 3}]
```

with the results shown in Table 15.4 at the top of the next page.

The probability for the possibility of two red balls and one black ball is given in the third row of the table as 0.1909. Verify this value by calculating the probability for a red ball on the first draw, a red ball on the second draw, and a black ball on the third draw according to the recipe for the *Pólya Urn Scheme*. Then multiply by the multiplicity factor $W(M) = 3$,

$$\frac{3}{10} \times \frac{4}{11} \times \frac{7}{12} \times 3 = 0.1909$$

Supplemental Exercise 15.3.10: Create similar code for calculating the probabilities dependent on data.

Solution to Supplemental Exercise 15.3.10:

In analogy to the above **priorPredictive[ ]** based on finding the probability for the future frequency counts, $P(M_1, M_2, \cdots, M_n)$, a similar looking function

Table 15.4: *The probability for all four future frequency counts in the* Pólya Urn Scheme *sums to 1.*

Possibility	*Frequency count*	*Ordering*	*Probability*
1	$3b, 0r$	b, b, b	0.3818
2	$2b, 1r$	b, b, r	0.3818
2		b, r, b	
2		r, b, b	
3	$1b, 2r$	b, r, r	0.1909
3		r, b, r	
3		r, r, b	
4	$0b, 3r$	r, r, r	0.0455
Sum			1.0000

`posteriorPredictive[`*arg1, arg2, arg3***`]`** with three arguments instead of two can be written to compute the posterior predictive formula. With this formula, the probability of future frequency counts can be computed when conditioned on the known data. As a matter of fact, this code has already appeared in Supplemental Exercise 14.2.4 with the name shortened.

It implements in straightforward fashion,

$$P(M_1, M_2, \cdots, M_n \mid N_1, N_2, \cdots, N_n) = W(M) \times \frac{\Gamma(N + \mathcal{A})}{\Gamma(M + N + \mathcal{A})} \times \frac{\prod_{i=1}^{n} \Gamma(M_i + N_i + \alpha_i)}{\prod_{i=1}^{n} \Gamma(N_i + \alpha_i)}$$

After evaluating,

```
Table[posteriorPredictive[{3 - i, i}, {7, 3}, {0, 0}], {i, 0, 3}]
```

a list of probabilities for all four future frequency counts,

```
{0.3818, 0.3818, 0.1909, 0.0455}
```

is produced, exactly the same probabilities found with **`priorPredictive[ ]`**.

The new second argument contains the data, here $N_1 = 7$ and $N_2 = 3$. The third argument is the alpha list of parameters supplied to the *beta distribution*, here $\alpha = 0$ and $\beta = 0$, simply to make the probabilities come out the same.

As mentioned, the same result could also have been achieved with,

```
Table[posteriorPredictive[{3 - i, i}, {6, 2}, {1, 1}], {i, 0, 3}]
```

Supplemental Exercise 15.3.11: Contrast this discussion of the Pólya Urn Scheme as an application of de Finetti's Theorem with Jaynes's explanation for the hypergeometric distribution for sampling without replacement and Jeffreys's formula for the same situation.

Solution to Supplemental Exercise 15.3.11:

Now, there is no way I can give a thorough and definitive response to such a question. That will have to wait for a later time. But I can solve a numerical example that perhaps points in the direction I am thinking.

In Chapter 3 of his book, Jaynes has an extensive discussion for the origin of the hypergeometric distribution in relation to sampling without replacement. He derives several alternative formats for the hypergeometric distribution, all of them interesting in their own right.

I will use the format presented in his Equation (3.74) (pg. 68) where he describes this formula as "the probability for drawing exactly r red balls and w white ones in $n = r + w$ draws from an urn containing R red and W white,"

$$h(r) = \frac{\binom{R}{r}\binom{W}{w}}{\binom{R+W}{r+w}}$$

Suppose, in consonance with our current running example, we were interested in the probability of drawing three black balls and two red balls without replacement from an urn containing a total of ten balls, seven of them black and the remaining three red.

Directly inserting these numbers into Jaynes's formula, we have,

$$\begin{aligned} P(M_1 = 3, M_2 = 2) &= \frac{\binom{B}{b}\binom{R}{r}}{\binom{B+R}{b+r}} \\ &= \frac{\binom{7}{3}\binom{3}{2}}{\binom{10}{5}} \end{aligned}$$

Now I will begin what seems like a rather unmotivated series of operations, but it

will become quite clear in the end what the goal is.

$$
\begin{aligned}
P(M_1 = 3, M_2 = 2) &= \frac{\frac{7\times6\times5\times4\times3}{4!}}{\frac{10\times9\times8\times7\times6}{5!}} \\
&= \frac{7\times6\times5\times4\times3}{4!} \times \frac{5!}{10\times9\times8\times7\times6} \\
&= \frac{7}{10}\times\frac{6}{9}\times\frac{5}{8}\times\frac{4}{7}\times\frac{3}{6}\times 5 \\
&= \frac{7}{10}\times\frac{6}{9}\times\frac{5}{8}\times\frac{3}{7}\times\frac{2}{6}\times\binom{5}{3}
\end{aligned}
$$

The last line is set up for $P(bbbrr)$ if we were to directly write out the probability for sampling the three black balls and then two red balls without replacement. The last term is $\binom{n}{k} = \binom{5}{3}$ for the multiplicity factor determining all the different ways for drawing three black balls and two red balls.

Immediately we notice the similarity to what we were doing before with the *Pólya Urn Scheme* and de Finetti's Theorem. Therefore, is it also possible to set up the probability for this sampling without replacement scenario, using as a template Feller's version of de Finetti's Theorem,

$$P\{S_n = k\} = \binom{n}{k}\int_0^1 \theta^k\,(1-\theta)^{n-k}\;F\{\mathrm{d}\theta\}$$

or, in my notation, as,

$$P(M_1, M_2) = \int_0^1 W(M)\;q^{M_1}\,(1-q)^{M_2}\;C_{\text{Beta}}\;\;q^{\alpha-1}\,(1-q)^{\beta-1}\,\mathrm{d}q?$$

What values for α and β return the correct probability?

Supplemental Exercise 15.3.12: Compute the above probability from the hypergeometric distribution.

Solution to Supplemental Exercise 15.3.12:

Before we attempt to answer the question we left hanging at the end of the last exercise, use *Mathematica's* built–in capability for calculating probabilities under the hypergeometric distribution.

`HypergeometricDistribution[`*arg1, arg2, arg3***`]`** demands three arguments. To begin, we shall follow the *Mathematica* documentation notation to describe these three arguments as they relate to our current inferential scenario.

The first argument n is the sample size being drawn from the urn. For our problem here $n = 5$ since we are drawing five balls from a total of ten in the urn.

The second argument n_{succ} is the total number of black balls in the urn and called the "number of successes." Here $n_{succ} = 7$. Obviously the "number of failures" would be identified with the number of red balls in the urn. The third and final argument n_{tot} is the total number of balls in the urn, here $n_{tot} = 10$. Thus, filling in these values we have an expression looking like,

```
HypergeometricDistribution[5, 7, 10]
```

Since we want a discrete set of numerical probabilities in our problem for any number of black and red balls drawn, we use **PDF[***dist, x***]** with two arguments, the first being our hypergeometric distribution as just set up, and the second argument the evaluation at x.

Let's now write,

```
Table[PDF[HypergeometricDistribution[5, 7, 10], k], {k, 0, 5}]
```

for calculating the six probabilities of zero through five black balls. For example, at $k = 2$ we are asking for the probability of two black balls and $n - k = 3$ red balls.

As an initial sanity check, we would require the first two probabilities to be zero. It is impossible to draw four or five red balls from the urn if there are only three to begin with. Remember, this is sampling *without* replacement.

The evaluation returns the list,

$$\{0, 0, \tfrac{1}{12}, \tfrac{5}{12}, \tfrac{5}{12}, \tfrac{1}{12}\}$$

with the easy translation that the probability of drawing two or five black balls is equal to 1/12, while drawing three or four black balls has the same probability of 5/12. Drawing three black balls means, of course, that two red balls were also drawn in the sample of five balls taken from the urn.

In Supplemental Exercise 15.3.11, we stopped right before the final computation of the probability for drawing three black balls because we wanted to highlight the explicit appearance of the probabilities at each draw,

$$P(M_1 = 3, M_2 = 2) = \frac{7}{10} \times \frac{6}{9} \times \frac{5}{8} \times \frac{3}{7} \times \frac{2}{6} \times \binom{5}{3}$$

The result confirms the hypergeometric distribution. The fourth element in the list of probabilities is,

$$\begin{aligned} P(M_1 = 3, M_2 = 2) &= 10 \times \frac{7}{10} \times \frac{6}{9} \times \frac{5}{8} \times \frac{3}{7} \times \frac{2}{6} \\ &= \frac{5}{12} \end{aligned}$$

Supplemental Exercise 15.3.13: Is it possible to discern Jaynes's formula within *Mathematica*?

Solution to Supplemental Exercise 15.3.13:

Yes, quite easily. If you choose to ask *Mathematica* for the symbolic representation of the hypergeometric distribution by substituting symbolic entries for the three arguments using the notation as described above in the previous exercise,

```
PDF[HypergeometricDistribution[n, nsucc, ntot], k]
```

the following expression is returned,

```
Binomial[nsucc, k] * Binomial[- nsucc + ntot, - k + n] /
                                        Binomial[ntot, n]
```

Here we have another opportunity to remark on *Mathematica*'s curious convention of reporting back expressions $n - k$ as **- k + n**.

Not surprisingly, the function **Binomial[***x, y***]** returns the binomial coefficient,

$$\binom{x}{y} \equiv \frac{x!}{y!\,(x-y)!}$$

The only issue is matching up the notation in Jaynes's formula,

$$h(r) = \frac{\binom{R}{r}\binom{W}{w}}{\binom{R+W}{r+w}}$$

with the above *Mathematica* expression.

It's easiest to just go ahead and substitute the numerical values we have been using in the exercises. With the following values of the arguments,

nsucc = 7, **k = 3**, **n = 5**, and **ntot = 10**

$$\binom{R}{r} \equiv \binom{B}{b} = \binom{7}{3} = \texttt{Binomial[7, 3]}$$

$$\binom{W}{w} \equiv \binom{R}{r} = \binom{3}{2} = \texttt{Binomial[- 7 + 10, - 3 + 5]}$$

$$\binom{R+W}{r+w} \equiv \binom{B+R}{b+r} = \binom{7+3}{3+2} = \texttt{Binomial[10, 5]}$$

In the end, we have the very satisfactory probability for drawing three black balls ($k = 3$) and two red balls ($n - k = 2$) as,

$P(k = 3) =$ **`Binomial[7, 3] * Binomial[3, 2] / Binomial[10, 5]`**

or as Jaynes wrote the hypergeometric probability for sampling without replacement,

$$h(r = 3) = \frac{\binom{7}{3}\binom{3}{2}}{\binom{10}{5}} = \frac{5}{12}$$

in Chapter 3, page 68, Eq(3.74).

This is the same formula for sampling without replacement as Jeffreys wrote it. See Chapter 6, page 118, 6.5.13.

Chapter 16

Predicting the Behavior of Cellular Automata?

16.1 Computational Irreducibility

The core notion behind probability and inferencing is that an IP can, in some sense, shortcut the actual details of an evolving Universe and *predict* future events. Opposing this hope is Wolfram's pessimistic warning encapsulated in his principles of computational irreducibility and undecidability.[1]

It's curious psychological fact that the way Wolfram's mind orders things is exactly the opposite of the way I would like to do it. For example, here is Wolfram in the referenced blog commenting on statistical mechanics:

> Like in an Ising model. Where one normally assumes that spins are kicked around by a heat bath. Well, there's a simple deterministic model that seems to have the same average behavior—and that's probably much closer to an actual spin system. It seems like there are a lot of probabilistic models where there are deterministic systems that have the same behavior.

I would prefer to find probabilistic models that mirror deterministic systems so that I can then leverage inferencing to predict the future behavior of the deterministic system without having to wait and watch it evolve.

Wolfram describes computational irreducibility as follows:

> Let's say you know the rules and initial conditions for a system. Well, then you can certainly work out what the system will do just by explicitly running

[1] Wolfram talks about computational irreducibility in many places. The first mention is in his book, *A New Kind of Science*, pp. 737–750. I have also found his blog post *A New Kind of Science and the Future of Mathematics* from 2004 to be particularly insightful on these topics as well.

> it. But the question is whether you can somehow shortcut that process. Can you for example just work out a formula for what will happen?
>
> That kind of computational reducibility is at the core of most traditional theoretical science. If you want to work out where an Earth in an elliptical orbit will be a million years from now, you don't have to trace a million orbits; you just have to plug a number into a formula.

We want to pose the question of what shading, or nuanced meaning, the concept of *computational irreducibility* assumes in the context of probabilistic inferencing? We do claim, as in Wolfram's explanation above, to be in possession of a formula that shortcuts the details of watching the detailed evolution of an ontological system far into the future. Unfortunately, inherent in the very nature of probability, with each further look into the future, the available probability must spread itself out over an ever increasing range of possibilities. In the end, combinatorial explosion does us in.

Supplemental Exercise 16.1.1: As an elementary illustration of the above remarks, revisit Jeffreys's animal with feathers inferential scenario.

Solution to Supplemental Exercise 16.1.1:

Jeffreys was stunned when he realized that after sampling half of some population, and finding that every single sample was one and the same, the probability was only 1/2 that the rest of the population was also the same as what was observed in every single sample so far!

In Exercise 44.7.9 of Volume III, we verified this probability of 1/2 by using the formula for the posterior predictive probability. As a numerical review, suppose that the population in question consists of 150,000 entities. 75,000 have been sampled and every single one of those 75,000 was found to be categorized as a Type I. What is the probability that the remaining 75,000 are also Type I as opposed to Type II?

With the data recorded as all Type I, $N_1 = 75,000$, $N_2 = 0$, and $N = 75,000$. Given the remaining unsampled population, the total number of future frequency counts is $M = 75,000$. The total size of the population is $M + N = 150,000$. We want the probability that all are Type I, $M_1 = 75,000$ and $M_2 = 0$. Plugging these numbers into the posterior predictive formula, we have,

$$\begin{aligned} P(M_1, M_2 \,|N_1, N_2) &= \frac{M!\,(N+n-1)!}{N_1!\,N_2!\,(M+N+n-1)!} \times \frac{(M_1+N_1)!\,(M_2+N_2)!}{M_1!\,M_2!} \\ &= \frac{75000!\,75001!}{75000!\,0!\,150001!} \times \frac{150000!\,0!}{75000!\,0!} \\ &= \frac{75001}{150001} \approx \frac{1}{2} \end{aligned}$$

But when we inquire about our degree of belief that the very next specimen from the unsampled population is a Type I, the same formula tells us that it is almost certain to be true,

$$P(M_1 = 1, M_2 = 0 \,|N_1, N_2) = \frac{75001}{75002} \approx 0.99999$$

The posterior predictive formula is indeed the correct formula that we would like to use in order to shortcut the computational irreducibility of actually sampling the next 75,000 specimens. It divides up the degree of belief in the sense of allocating 1/2 for seeing 75,000 of Type I, and 1/2 of not seeing 75,000 of Type I. In other words, the left over 1/2 is divided up into the probability of seeing 74,999 of Type I and one of Type II, or seeing 74,998 of Type I and two of Type II, or seeing 74,997 of Type I and three of Type II, $\cdots$, or all 75,000 of Type II.

I don't find it that shocking that out of the next 75,000 entities that remain in the population, there might be a few Type I exceptions. As a matter of fact, the probability of seeing one, two, or three Type IIs is already at 0.4375. One might just as reasonably be shocked in the opposite sense of Jeffreys by the prospect of not seeing a couple of Type IIs in the remaining 75,000 even if the first 75,000 were all Type Is.

As mentioned in Volume III, the details of how that first half of the population was sampled is never completely specified. So there is always some rationale that the sampling methodology was following its own restricted logic, and just hasn't yet encountered that unique set of circumstances which in the second half of the population will yield up an exception.

As I whimsically suggested, when Sir Harold was sampling the population of birds and their possession of beaks, he hadn't yet got around to sampling outside of Cambridge, or outside of England, or outside of Europe, or outside of this planet, or outside of the evolutionary quirks of this geological epoch.

16.2 One Model is Deduced

Suppose an IP is trying to adhere closely to our recommendations for a posterior prediction of the color of the cells in a cellular automaton. An unknown CA has been observed with some given initial conditions for sixteen cells. The only data are the colors of the sixteen cells after the evolution of one time step.

After examining the data, the IP deduces that the model must be Rule 110. With the luxury of only a single model to contend with, the IP can confidently predict the color of the sixteen cells at the next time step. As another example of a probability expression and an inference generalizing a deduction, write out the posterior predictive probability for the color of the cells,

$$P(B_{N+1}, C_{N+1}, \cdots \,|\, A_N, B_N, C_N, D_N, \cdots, \mathcal{D}) =$$

$$\sum_{k=1}^{\mathcal{M}} P(B_{N+1}, C_{N+1}, \cdots, \,|\, A_N, B_N, C_N, D_N, \cdots, \mathcal{M}_k)\; P(\mathcal{M}_k \,|\, \mathcal{D})$$

Since there is only one model, Rule 110, at the second time step $N + 1 = 3$.

$$P(B_3, C_3, \overline{D}_3 \,|\, A_2, \overline{B}_2, C_2, D_2, E_2, \mathcal{D}) = P(B_3, C_3, \overline{D}_3 \,|\, A_2, \overline{B}_2, C_2, D_2, E_2, \text{Rule 110})$$

Making use of conditional independence,

$$P(B_3, C_3, \overline{D}_3 \,|\, A_2, \overline{B}_2, C_2, D_2, E_2, \text{Rule 110}) =$$

$$P(B_3 \,|\, A_2, \overline{B}_2, C_2, \mathcal{M}_k) \times P(C_3 \,|\, \overline{B}_2, C_2, D_2, \mathcal{M}_k) \times P(\overline{D}_3 \,|\, C_2, D_2, E_2, \mathcal{M}_k) = 1$$

The IP is certain that the colors of cells 2, 3, and 4 at step 2 will be black, black, and white given that the colors of the relevant five cells at the previous time step were black, white, black, black, and black,

$$P(b, b, w, \,|\, b, w, b, b, b, \text{Rule 110}) = 1$$

Supplemental Exercise 16.2.1: Draw a diagram of a cellular automaton operating according to Rule 110 with given initial conditions and evolving for just two time steps.

Solution to Supplemental Exercise 16.2.1:

Figure 16.1 at the top of the next page illustrates all of this. The *Mathematica* code for reproducing this picture is,

```
ArrayPlot[CellularAutomaton[110,
  {1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0}, 1], Mesh → True]
```

The first argument to **CellularAutomaton[ ]** is the rule number running the evolution of the system. The second argument is the list of initial conditions, the starting configuration of black and white cells, while the third argument specifies how many steps to evolve the CA from the initial conditions.

The diagram in the middle of Figure 16.1 showing the eight possibilities for the colors of the three relevant cells that will determine the color of the cell to be updated by Rule 110 is generated by,

```
RulePlot[CellularAutomaton[110]]
```

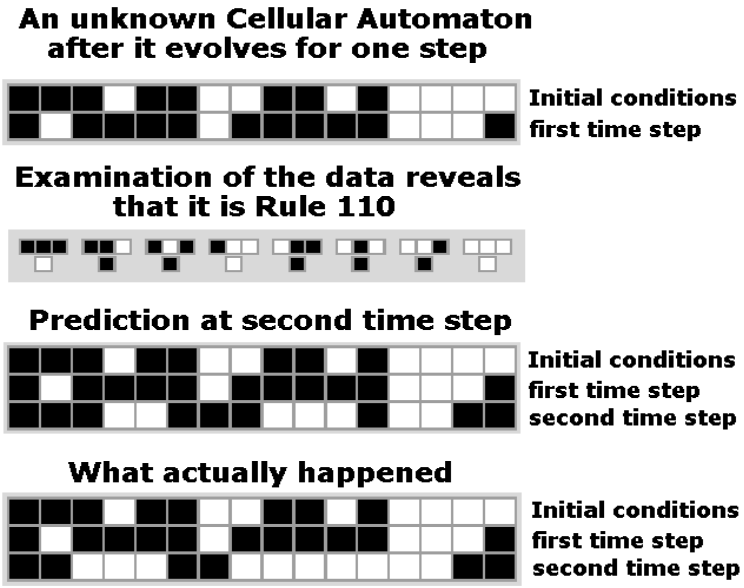

Figure 16.1: *Examining the data from an unknown CA reveals that it must be running according to Rule 110. However, the prediction at the next time step is not correct.*

The colors of all sixteen cells in the third row of Figure 16.1 are shown by letting the system evolve one more step under Rule 110. This, of course, just mandates a change in the third argument from 1 to 2. This is the confident prediction made by the IP whenever the system can actually be observed at the second time step.

An unforeseen catastrophe has struck! The colors of the cells at the next time step are not what was predicted. The IP has made the mistake of assuming that the ontological system was restricted to a logic function with three arguments. Suppose however that the Universe chooses to be more complicated than the IP initially assumed, and instead relies on a logic function with five arguments to evolve the system.

There are $2^{2^3} = 256$ logic functions (the 256 ECA Rules) with three arguments and $2^{2^5} = 4,294,967,296$ logic functions with five arguments. The CA still updates the colors of its cells with only two colors, black and white, but now looks at the colors of the *two* neighboring cells to the left, and the *two* neighboring cells to the right of the cell above the cell to be updated. Previously, Rule 110, as with all of the 256 rules based on logic functions with three arguments, looked at just one neighboring cell on the left and right.

As will be demonstrated, Rule 743,977,160, selected from the $4,294,967,296$ available rules for two neighbors on each side and two colors, starting out from the same initial configuration as before, not only matched the data at step 1, but the data at step 2 as well.

Supplemental Exercise 16.2.2: Go through all of the details in figuring out what rule generated the cellular automaton matching the observed data over the first two time steps.

Solution to Supplemental Exercise 16.2.2:

The actual data observed at the first and second time steps are shown at the bottom of Figure 16.1. It is clear that the predictions made by Rule 110 for time step 2 did not come to pass, even though the rule matched all of the data at the first time step.

However, Rule 743,977,160 does manage to match all of the data seen at both time steps. Examining the output for the first two time steps confirms this,

```
ArrayPlot[CellularAutomaton[{743 977 160, 2, 2},
  {1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0}, 2], Mesh → True]
```

The details involve examining the digits of the base 2 expansion of a rule number. We know that there will be 32 digits, either 0 or 1, needed to express a number between 0 and $4,294,967,295$ as $2^{31} + 2^{30} + \cdots + 2^0$.

The built–in function **IntegerDigits[ ]** will provide us with a correct list of 32 1s when we inspect the last rule number,

```
IntegerDigits[4 294 967 295, 2, 32]
```

and a list with 32 elements starting and ending with a 1 and all other elements equal to 0 when we inspect a middle of the pack rule number,

```
IntegerDigits[2 147 483 649, 2, 32]
```

Recall that this was the rule used in section 3.6.1 and Exercise 3.7.10 of Volume I, and most recently revisited in Supplemental Exercise 3.2.7. It is an elementary illustration of a cellular automaton operating according to a rule that is the same as a logic function with five arguments.

But we would really like to have the inverse of this function where we could input a list of 32 1s and 0s and get in return the decimal number. *Mathematica* does not fail us here; it provides the built–in function **FromDigits[ ]** where the first argument is a list of digits, and the second argument is the base. So, for this example,

$$\texttt{FromDigits[\{1, } \overbrace{0, 0, \cdots}^{30}\texttt{, 1\}, 2]}$$

returns 2,147,483,649.

This is where the tedious part of the exercise begins. We want to find a rule number based on the black and white colors of five cells that matches the already

seen output from the initial conditions given for Rule 110. Examining Figure 16.1, and scanning from left to right for the first desired output of a black cell in the third cell at the first time step, we require that some rule output a black cell for the pattern of b, b, b, w, b, namely, the first five cells in the initial conditions. But this is the same as finding the unknown logic function that outputs a T with the five arguments of $f_?(T, T, T, F, T) = T$.

Generate a list of lists with **`Tuples[{T, F}, 5]`** showing the five arguments in *Mathematica*'s preferred order. The first list would be **`{T, T, T, T, T}`**, and the last list would be **`{F, F, F, F, F}`**. There are 32 lists like these collected into the outer list. Let's make it easier to pick out the order number of any particular list of arguments with,

```
Column[Table[Row[{i, Tuples[{T, F}, 5][[i]]}], {i, 32}]]
```

We now scan this layout to find where **`{T, T, T, F, T}`** occurs. It happens to be the third element. The function must have a functional value of T for these five arguments.

Thus, in order to begin constructing the list of digits as the first argument to **`FromDigits[ ]`**, we have found out that a "1" must appear in the third place. In other words, the digit 1 will appear as the contribution from 2^{29} in the overall determination of the rule number in decimal format. Wolfram orders the colors for a rule number starting with all black and ending with all white. So, if the argument **`{T, T, T, T, T}`** were to be assigned a T, then the placement of the digit 1 would begin at the left end with 2^{31}.

Continue on in the same manner. The fourth cell at the first time step was also black. The relevant cell colors set by the initial conditions were now, moving over one place, b, b, w, b, b. This corresponds to **`{T, T, F, T, T}`**. This is the fifth element from **`Tuples[ ]`**. Since the updated cell is black, the functional assignment must once again be a T. The digit "1" must be inserted at the fifth place as our construction of the list of "0"s and "1"s continues.

As an example of where the digit 0 would be inserted into the growing list as the argument for **`FromDigits[ ]`**, note that the data told us that the seventh cell was colored white. The pattern of cells in the initial conditions that produced this white cell was b, b, w, w, b. **`{T, T, F, F, T}`** occurs in the seventh position, so a "0" must be inserted at the seventh position. Summarizing the construction of our list to this point, there must be a "1" in third position, a "1" in the fifth position, and a "0" in the seventh position.

As I said, this is all very tedious. But eventually, the following list of 32 "1"s and "0"s is built up,

```
{0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,
 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0}
```

and now it is easy sledding to find the appropriate rule number with,

```
FromDigits[{0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,
            0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0}, 2]
```

returning rule number 743,977,160 as the logic function which will duplicate all of the observed data seen so far.

Supplemental Exercise 16.2.3: Where does the conceptual error occur for the above inference of predicting the colors of the cells in a cellular automaton?

Solution to Supplemental Exercise 16.2.3:

At first glance, our tendency is to blame the IP for specifying an inadequate model space. We criticize it for stopping too early in its deliberations of what class of rules is governing the ontological system. But further thought reveals that it is a fundamental inability in specifying the proper state space that is the real root of the problem.

For the ECA, the state space was correctly set up for statements involving A_N, B_N, C_N and B_{N+1}. There was no problem in conceptualizing the numerical assignments to the sixteen cells of a joint probability table under any model.

Likewise, for a CA running under a more complicated rule, as in the above exercise, we could expand the state space to A_N, B_N, C_N, D_N, E_N and C_{N+1}. Any model can theoretically fill in all 64 cells of a joint probability table required for this inference.

It is our fundamental ignorance about what elements should be in the state space that describes how the Universe ticks over that really stymies an IP. Once a correct state space has been established, there is no conceptual problem in thinking about all of the models making assignments to the cells of the joint probability table.

We take away an important and sobering lesson from the last exercise. An IP never really knows for sure whether its tentatively entertained state space is adequate to the task. Even after expanding to a more complicated rule, we still aren't completely sure whether the data at the third time step might not deviate from the prediction of Rule 743,977,160. And then we would have to start all over again trying to discover a yet even more complicated rule.

Everyone seems to agree that the most well–known example to date for the manner in which the scientific enterprise lumbers forward, as I have tried to illustrate abstractly through ontological systems like cellular automata, centers on an initial universal adoption of Newton's Law of Gravitation. It seemed to fit all of the data observed until a few anomalies started cropping up. One of the most notorious of these anomalies with Newton's Law was a discrepancy in the orbit of the planet Mercury. It took an Einstein to replace a Rule 110 with a Rule 743,977,160 so to speak.

But this iterative procedure of tentatively adopting, critiquing, and then revising the state space as anomalies occur, is crucial to the scientific enterprise. It shouldn't be judged too harshly; it's the best we can do with the tools given us.

16.3 Searching for New Physics

Let's do what Einstein did for Newton. After discovering an unsuspected anomaly in our ontological system, search for a revision to our existing Physics. After "fixing" the anomaly, we will tentatively assume that this new Physics governs our world until the next mishap.

Of course, I'm being a bit cheeky here for fun. But there is a serious side to this exercise. For the purpose of the *GedankenExperiment*, our Universe is a cellular automaton, an abstract ontological system that is computable.

We originally thought the Physics of our Universe was run by Rule 110 after fitting the data at the first time step. After discovering that Rule 110's prediction at time step 2 did not coincide with the way our world works, we had to update it with new Physics to Rule 743,977,160. The new Physics took the revolutionary step of declaring that the color of a cell depended not on three cells as we originally surmised, but in what was a scientific surprise to us, actually depended on five cells.

Indulging in the iterative scientific enterprise as the only way we know how to proceed, we once again confidently predict the world's behavior from our newly revamped Physics. Should it happen that another anomaly occurs in our cellular automaton, we are forced back to the starting blocks searching for a yet another more complicated rule that explains everything that has happened so far, just as Rule 110 and Rule 743,977,160 did for a while, but in addition has to explain away the new anomaly.

I intend for this exercise to be a way of expressing the difference in the effort at changing some piece of underlying scientific knowledge versus the use of probability theory to conduct inferencing.

Supplemental Exercise 16.3.1: What changes to the currently accepted state space would have to be undertaken if the anomaly of a *gray* cell happened to be observed as data in the universe governed by our cellular automaton?

Solution to Supplemental Exercise 16.3.1:

The bottom half of Figure 16.2 shows what happens in our World at the third time step. Contrary to the prediction of the "Old Physics," the completely new phenomenon of a gray cell occurs at cells 10, 11, and 12.

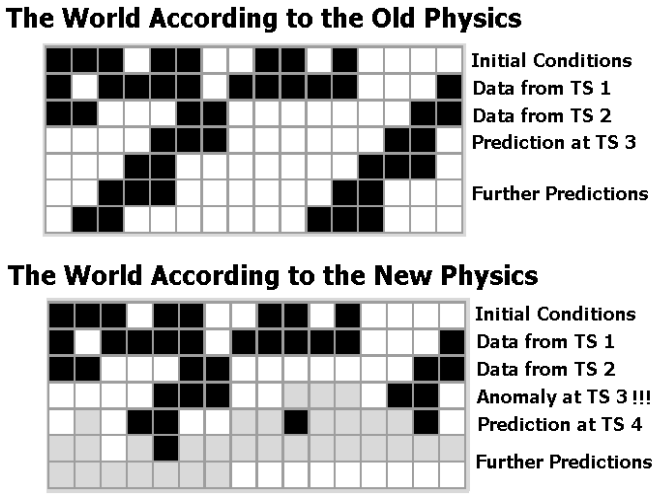

Figure 16.2: *Examining the initial data about our World suggested that it might be running according to Rule 110. Further data demanded revision of this first tentative "Physics." As data about our World continue to accumulate, further revisions to the tentatively accepted "Physics" are required.*

Taxing the limits of our concentration, but fortunately for us not *Mathematica*'s ability to compute cellular automata, examine the New Physics of a three color, five variable cellular automaton to explain away the unexpected anomaly of a gray cell appearing as data at the third time step. Our mode of attack is the same as in the last supplemental exercise.

In our attempt to discover the "New Physics" describing all of the data seen so far, we move up from a state space with only two colors possible for a cell to allowing three colors for each cell. Trying the best we can to follow Ockham's dictum, we retain dependency on the colors of the five relevant cells at the previous time step.

Even this early on in the game, combinatorial explosion flouts our efforts with a vengeance. Wolfram's formula for the total number of rules for three colors, $k = 3$, and two neighbors on the left and right side of the cell above the cell to be updated, $r = 2$, tells us that there are only,

$$k^{k^{(2r+1)}} = 3^{3^5} \approx 8.72 \times 10^{115}$$

number of possible rules, a number with 116 decimal digits.

With the introduction of the possibility of a gray colored cell, we must move up from the binary number system to a number system with base 3. So, any rule number from the astronomically huge number of decimal rules will consist of $3^5 = 243$ digits, where the digits can only be 0, 1, or 2.

In the previous supplemental exercise, we examined a list with $2^5 = 32$ elements consisting solely of the digits 0 and 1. We now have to face up to a list of 243

elements consisting solely of the digits 0, 1, or 2 in order to represent a rule as an argument to **`CellularAutomaton[ ]`**.

This rule number will be expanded in base 3 as,

$$\text{Decimal number} = (d \times 3^{242}) + (d \times 3^{241}) + \cdots + (d \times 3^{0})$$

where d is 0, 1, or 2.

Selecting some easy first examples, we have for rule 7 and rule 173,

$$7_{10} = 21_3$$

$$173_{10} = 20102_3$$

and thus a list of 243 elements with the first 241 all **`0`** followed by a **`2`** and a **`1`** for rule 7, $\{\overbrace{\cdots}^{241\ \mathbf{0}\text{s}}, 2, 1\}$, and a list of 243 elements with the first 238 all **`0`** followed by **`2`**, **`0`**, **`1`**, **`0`**, and **`2`** for rule 173, $\{\overbrace{\cdots}^{238\ \mathbf{0}\text{s}}, 2, 0, 1, 0, 2\}$.

We can construct any list of 243 elements just like this and then use,

`FromDigits[{`$\cdots$**`}, 3]`**

to find out what the decimal rule number will be. For example,

`FromDigits[{`$\overbrace{\cdots}^{237\ \mathbf{0}\text{s}}$**`, 2, 0, 1, 2, 2, 1}, 3]`**

processes the list of 243 digits and returns the decimal rule number 538 as a possible first argument to **`CellularAutomaton[ ]`**.

We would want to keep the representation of any rule number in this form as a list of 243 0s, 1s, or 2s. For example, if just the first element in the list is a "2" followed by 242 "0"s, the resulting decimal rule number is one of those astronomical numbers consisting of 116 decimal digits.

The expansion in base 3 for this list of digits would be 2×3^{242}. Asking *Mathematica* to compute **`Times[2, Power[3, 242]]`** results in rule number 58, $\overbrace{\cdots}^{108\ \text{digits}}, 110, 418$.

Keeping the rule number as a list of 0, 1, and 2 is absolutely vital for performing "genetic engineering" on the list of digits. The list representation allows the IP to surgically snip out and replace whatever digits are necessary to modify an existing rule number. This process takes place in order to satisfy the need to match all of the observed data.

For example, here is a little utilitarian function I wrote to help me watch a cellular automaton iteratively attempt to match the known data of our World at the first two time steps.

```
ap[oldList_List, k_, j_, steps_] := Module[{dummy},
   newList = ReplacePart[oldList, 244 - k → j];
   ArrayPlot[CellularAutomaton[{FromDigits[newList, 3], 3, 2},
   {1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0}, steps],
   ColorRules → {0 → White, 1 → Black, 2 → LightGray},
                                          Mesh → True]]
```

The two global variables **oldList** and **newList** were initialized as a list of 243 2s and a list of 243 0s, respectively. Running **ap[oldList, 2, 0, 6]** allowed me to replace the digit 2 at the second position with the digit 0 to see what effect this had on trying to match the data at the first time step. When this CA produced a white colored cell at the next to the rightmost cell at the first time step, I knew I was on the right track.

In the same manner, surgically replacing the existing digits iteratively with a digit appropriate to match the data permitted me to build up to the final cellular automaton that matched all of the existing data. This was the culmination of the development of the "New Physics" which then predicted the colors of all the cells for the fourth time step and beyond. Obviously, these predictions from the New Physics differs from the predictions made by all of the previous versions of the Old Physics.

There are a few more tedious, but vital, details to flesh out before we can call an end to this exercise. Insertion of a particular digit in the digit list specified what color the cell would assume. As observed from the above *Mathematica* program, a "0" specified a white cell, a "1" a black cell, and a "2" a gray cell.

But what pattern of five cells did each digit in the digit list refer back to? Specifying the last element in the digit list as a "1" for example means that some cell will be colored black. Specifying the first element in the digit list to be a "2" means that some cell will be colored gray. What is required is some means of matching up the construction of the digit list with the particular pattern of the relevant five cells at the previous time step.

The answer to this question is to use **Tuples[{W, B, G}, 5]** to find out how Wolfram wants to order all 243 combinations of five white, black, and gray cells. Put things into a table so we can scan down the list and pick out which 5–tuple we are interested in,

```
Column[Table[Row[{i, Spacer[10], Tuples[{W, B, G}, 5][[i]]}],
                                                    {i, 243}]]
```

We see that the first item is a list of all five white cells **{W, W, W, W, W}**, the thirtieth item is the list **{W, B, W, W, G}**, and so on through all 243 possibilities, until we arrive at the final item consisting of all five gray cells, **{G, G, G, G, G}**.

A number expansion in a particular base is just like a function expansion with coefficients and orthogonal building block functions. We first used this concept in Volume I, Chapter Four, where we wrote down the orthonormal expansion of a

Boolean function with two arguments as,

$$f(A,B) = \sum_{i=1}^{4} c_i(A,B)\,\phi_i(A,B)$$

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

Each one of the 243 five–tuples matches the corresponding orthogonal building block functions 3^0, 3^1, $\cdots$, 3^{242}. The digit of 0, 1, or 2 is the coefficient for each 3^0, 3^1, $\cdots$, 3^{242}.

Thus, the first item in our above listing, **{W, W, W, W, W}**, corresponds to the orthogonal building block function 3^0. If we specify the coefficient as say the digit "1" then this is Rule 1, telling the cellular automaton to output a black cell when the previous cell and its two neighbors to the left and right were all white. All other configurations result in a white cell. The second item in our above listing, **{W, W, W, W, B}**, corresponds to the orthogonal building block function 3^1. If we specify the coefficient as say the digit "2", then this is Rule 7, telling the cellular automaton to output a gray cell when the previous cell was white, its two neighbors to the left were white, and its two neighbors to right were white and black. We obviously retained the coefficient "1" from the previous example to obtain Rule 7.

The thirtieth item in our above listing **{W, B, W, W, G}** corresponds to the orthogonal building block function 3^{29}. If we specify the coefficient as say the digit "0" then this is Rule $(0\times 3^{29})+\cdots+(2\times 3^1)+(1\times 3^0)$, telling the cellular automaton to output a white cell when the previous cell was white, its two neighbors to the left were white and black, and its two neighbors to right were white and gray.

The final item, in other words, the final five-tuple, in the listing is the 243rd item, **{G, G, G, G, G}**. This corresponds to the final orthogonal building block function 3^{242}. If, in order to match some existing data that showed five gray cells colluded to produce a black cell at the next time step, the digit "1" would be specified as the coefficient. The rule number would be $(1\times 3^{242})+\cdots+(0\times 3^{29})+\cdots+(2\times 3^1)+(1\times 3^0)$ implementing this as well as all of the other specifications of an updated cell color given the five previous relevant variables.

Using **ap[ ]**, the initial digit list of all 2s (it could have been all 0s, or all 1s, or a random selection of digits) was genetically engineered by iteratively snipping and splicing at the right locations to match all of the existing data. For example, looking at the data at the first time step, we see that the initial conditions existing for the leftmost five cells of b, b, b, w, b have produced a black cell in cell 3. Scanning the list of **Tuples[ ]** reveals that this is at location 119. Replace the already existing digit 2 in the digit list with a 1 at this location. Continue to modify the current digit list by matching the rest of the data until finished.

The rule number,

$$29,\ \overbrace{\cdots}^{108\text{ digits}},867,180$$

resulting from this surgically altered list of digits was entered as the first argument to **`CellularAutomaton[ ]`**. It produced the picture of the "New Physics" at the bottom of Figure 16.2.

Supplemental Exercise 16.3.2: Insert a random digit list as the first argument to `CellularAutomaton[ ]`.

Solution to Supplemental Exercise 16.3.2:

`RandomInteger[2, 243]` will produce a list with 243 randomly selected digits consisting of 0, 1, or 2. Then, **`FromDigits[RandomInteger[2, 243], 3]`** will produce the decimal rule number for some possible three color, five variable CA. Evolve this CA for, say, 200 steps to ascertain the presence of any of the five properties listed at the top of the next page,

```
ArrayPlot[CellularAutomaton[
            {FromDigits[RandomInteger[2, 243], 3], 3, 2},
            {1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0}, 200],
         ColorRules → {0 → White, 1 → Black, 2 → LightGray},
                        Mesh → True, PixelConstrained → 10]
```

Inspect the beginning of the CA shown in Figure 16.3 below,

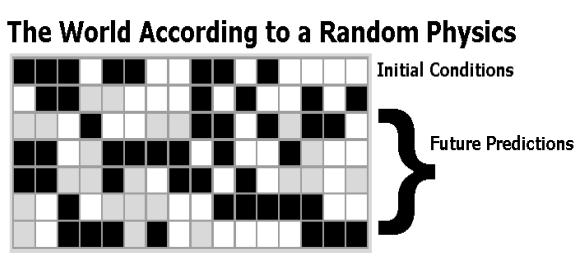

Figure 16.3: *An ontological system represented as a CA evolving according to some random rule.*

The initial conditions were the same, but this World evolved according to a different Physics. The fourth cell over from the left at the first time step is colored gray. The colors of the five relevant cells in the initial conditions determining this color were b, b, w, b, b. Scanning the list produced by **`Tuples[ ]`**, the digit "2" for gray matches up with the 113th orthogonal building block function in the number expansion. Whatever the expansion of the random rule number in base 3 turns out to be, and we could easily find out what, in fact, it is with **`FromDigits[ ]`**, there must be a term 2×3^{112} contributing to the total in the number expansion.

Here are some of the physics questions we would like to explore by examining this toy ontological World:

Property 1: Does this World reach a state of complete stasis where nothing further happens? (all gray cells, say, at some future time step),

Property 2: Does this World oscillate among configuration of cells experienced before, and what is the periodic nature of the oscillation (repetitive and nested structures),

Property 3: Do neither of the above take place, and the World continues to evolve,

Property 4: If it does seem to evolve toward a non–equilibrium state, is it an *undecidable* evolution beset by *computational irreducibility*?, or,

Property 5: Are some macrostructures present in the far future of the World *predictable* through the intervention of inferencing and probability theory?

16.4 Probability Predictions

We mentioned earlier that we wanted to compare the general approach taken by CA as detailed in the last few exercises with the general approach we would employ if we were viewing things through an inferential prism. When dealing with CA, we thought in terms of predicting the colors of cells at the next time step $N+1$, given the colors of some number of relevant cells at the previous time step N.

To start off using the inferential approach, it is easier, I believe, to say that we are going to predict the outcome of some statement $(A_{N+1} = a_i)$ when conditioned on some number of predictor variables, $B_{N+1}, C_{N+1}, \cdots$, and the known data $\mathcal{D}$. In other words, the statement to be predicted together with all of its predictor variables are lumped together at time $N+1$. The data $\mathcal{D}$ is then neatly segregated notationally as $(A_1, B_1, C_1, \cdots, A_N, B_N, C_N)$.

The particular case that we have just been examining from the perspective of CA consisted of five predictor variables and three possible measurements. From the perspective of a general inference, the ultimate goal is an attempt to compute a conditional probability,

$$P(A_{N+1} \,|\, B_{N+1}, C_{N+1}, \cdots, F_{N+1}, \mathcal{D})$$

via Bayes's theorem.

Since $(A = a_i)$ can be in one of three states–white, black, or gray–the denominator in Bayes's Theorem will consist of three joint probabilities.

Supplemental Exercise 16.4.1: Given the above scenario mimicking the evolution of a CA, compute the posterior predictive probability.

Solution to Supplemental Exercise 16.4.1:

For a total of six variables each with the possibility of being observed in one of three conditions, the state space has a dimension of $n = 3^6 = 729$. Any specific joint probability table or contingency table would then be constructed with 729 cells.

But the IP can short circuit much of this labor by realizing that the posterior predictive probability averages over *all* models. There is no requirement, at least for the purposes of a posterior predictive probability, to explicitly fill in a 729 cell joint probability table through a specific model. Contrast this procedure with what was done for the CA. There, we tried very hard to find just *one* model, one rule number, that fit all of the known data in order to make our predictions.

It was an extremely pleasant surprise to find that the derivation of the posterior predictive probability, when averaged over all models and starting from Laplace's prescription for a complete lack of knowledge about causes, resulted in an almost embarrassingly simple formula. See Supplemental Exercise 14.2.3 for a review.

The joint posterior probability for any one of the 729 statements, at the very next trial,

$$P(A_{N+1}, B_{N+1}, \cdots, F_{N+1}, \mathcal{D})$$

is found from the formula developed for the probability of any number of future frequency counts conditioned on the known data of the past frequency counts,

$$P(M_1, M_2, \cdots, M_n \mid N_1, N_2, \cdots, N_n)$$

When $M = 1$, with the consequence that only one $M_i = 1$, and all the remaining $M_i = 0$, the formula simplifies to,

$$P(A_{N+1}, B_{N+1}, \cdots, F_{N+1}, \mathcal{D}) = \frac{N_i + 1}{N + n}$$

After an application of Bayes's Theorem,

$$P(A_{N+1}|B_{N+1}, \cdots, F_{N+1}, \mathcal{D}) = \frac{N_j + 1}{\sum_{j=1}^{n_A} N_j + 1}$$

where $n_A = 3$.

The hardest part of this whole enterprise is to develop some sort of data indexing scheme for how the observations will be ordered within the contingency table. Suppose that the $(A = a_j)$ statement is placed at the highest level so that all of the joint statements $(A = a_1, B = b_j, \cdots, F = f_j)$ occupy cells 1 through 243, the joint statements $(A = a_2, B = b_j, \cdots, F = f_j)$ occupy cells 244 through 486, and the joint statements $(A = a_3, B = b_j, \cdots, F = f_j)$ occupy cells 487 through 729.

For example, cell 3 will contain the past frequency counts, in other words, the data, for any occurrences of the statement $(A = a_1, B = b_3, C = c_1, \cdots, F = f_1)$, cell 246 will contain the past frequency counts, again, just more data, for any occurrences of the statement $(A = a_2, B = b_3, C = c_1, \cdots, F = f_1)$, and finally, cell 489 will contain the past frequency counts, the final bit of data, for any occurrences of the statement $(A = a_3, B = b_3, C = c_1, \cdots, F = f_1)$.

To compute the posterior predictive probability for A measured as a_2 conditioned on the status of the predictor variables and the known data, the formula might return a probability of 1/3, as the example below shows.

$$P(A_{N+1} = a_2 \,|\, B_{N+1} = b_3, C_{N+1} = c_1, D_{N+1} = d_1, E_{N+1} = e_1, F_{N+1} = f_1, \mathcal{D}) =$$

$$\frac{N_{246} + 1}{(N_3 + 1) + (N_{246} + 1) + (N_{489} + 1)} =$$

$$\frac{3}{4 + 3 + 2} = \frac{1}{3}$$

When the data base was probed, we found the past frequency counts of $N_3 = 3$, $N_{246} = 2$, and $N_{489} = 1$.

An inference about the degree of belief in the truth of the statement $(A_{N+1} = a_2)$ is analogous to predicting that a CA will produce a black cell given the disposition of the colors of the five relevant cells at the previous time step, and all of the previous observations of the CA at work.

The difference is between a certain prediction based on tentatively entertaining just one model for a CA, and an equivocable prediction based on all models for an inference. For the CA, we must discover some "New Physics" to explain away any surprising anomaly. Inferencing is continuing to do the same thing, but has whittled away at the relative degree of belief in all of the models representing all of the potential Physics whenever any surprising anomalies crop up.

Supplemental Exercise 16.4.2: What is the harder inferential problem?

Solution to Supplemental Exercise 16.4.2:

The IP would not, it seems, be so much interested in *microstatements* such as the color of one particular cell at one particular time step in the future, but rather would

want to assess the degree of belief in any "localized structures" that manifested themselves during the evolution of the CA representing our toy ontological World.

From the outset, we are faced with a different and more difficult inferential problem because the original state space would have to be re–defined. What is an appropriate state space for "localized structures?"

Jaynes had some insightful comments on the difficulty and importance of defining an initial state space in real world Physics. His most cogent insight was his advice to adopt a flexible attitude about the inherent fluidity of state spaces. What might work well in one problem might not be a universal answer to all problems. And even if the IP does manage to achieve some sort of plausible solution to this problem, given increased knowledge of the physical world, how does one update and redefine this initial state space?

Going all the way to the bottom by setting up a quantum version of the state space may not in itself be a good resolution. The analogy might be that we end up reverting back to where we began with predicting individual cell colors, miss the forest for the trees by not focusing on any "localized structures" which are the real explanatory features of the physics of the toy world.

Appendix A

Deconstructing *Mathematica* Code for Cellular Automata

A.1 The Motivation

I was perusing the on–line tutorial for cellular automata in the Wolfram Language Reference documentation when I came upon this statement together with some accompanying *Mathematica* code. I have no way of knowing for sure, but I have a hunch that this tutorial was written by Stephen Wolfram.

> Any $k = 2$ cellular automaton rule can be thought of as corresponding to a Boolean function. In the simplest case, basic Boolean functions like `And` or `Nor` take two arguments. These are conveniently specified in a cellular automaton rule as being at offsets `{{0}, {1}}`. Note that for compatability with handling higher–dimensional cellular automata, offsets must always be given in lists, even for one dimensional cellular automata.
>
> This generates the truth table for 2–cell–neighborhood rule number 7, which turns out to be the Boolean function `Nand`.
>
> ```
> Map[CellularAutomaton[{7, 2, {{0}, {1}}}, #, 1][[2, 2]]&,
                        {{1, 1}, {{1, 0}, {{0, 1}, {{0, 0}}]
> ```

It wasn't immediately, or for that matter a long time, after pondering this code, that I understood what this code was doing. This lack of apprehension on my part was even more frustrating because I had just spent a fair amount of time in Chapter Sixteen of these Supplemental Exercises writing out very similar looking expressions for cellular automata.

And, of course, many many words had been expended, both in Volume I and in these Supplemental Exercises, about the fundamental sixteen logic functions which take two arguments. For example, the `Nand` logic function has been looked at here in the context of how Jaynes and Garrett saw its relevance.

I had devoted whole Chapters attempting to show how cellular automata were invoking logic functions when they updated a cell. So you might understand the frustration over my obtuseness concerning this piece of code.

This Appendix also gives me an opportunity to express my difference of opinion with Wolfram over how transparent *Mathematica* is as a programming language. Now, let there be no doubt; I am a loyal, ardent, and devoted proselytizer for *Mathematica*. I am not going to learn any more programming languages.

Nevertheless, one will often end up spending a considerable amount of time in deconstructing even what seems to be a small amount of *Mathematica* code. It can be as opaque as the originator wishes, and any hoped for advantage in some sort of magical transparency of the language disappears down the drain. Immediate apprehension of even a short amount of code, to wit, many of Wolfram's examples in the Notes to *A New Kind of Science*, can be quite elusive.

A.2 The Goal

I will deconstruct, in some level of detail, the following *Mathematica* code used to generate cellular automata that output a truth table for some Boolean function,

```
CellularAutomaton[{7, 2, {{0}, {1}}}, #, 1][[2, 2]]&
                        /@ {{1, 1}, {{1, 0}, {{0, 1}, {{0, 0}}
```

and, at the end, add my own extra layers of non–transparency.

I thought that this exercise would be of some general interest since it affords us the opportunity to review several popular topics like logic functions, truth tables, cellular automata, and, most especially, some subtleties of *Mathematica* syntax.

A.3 The Deconstruction (the easier part)

A.3.1 Truth tables

A truth table computes the functional assignments for all of the possibilities the arguments to the logic function can take on.[1] I had introduced a function called **TruthTable[f_]** in Appendix A of Volume I for this task. I later mentioned that this effort, as valuable as it might have been, was superseded by the *Mathematica* built–in function called **Boolean Table[** *Boolean function* **]**.

The code we are eventually going to deconstruct will indeed generate the truth table for any logic function, but it is not the easiest way. The aforementioned **BooleanTable[ ]** is far more direct. For the particular elementary logic function **Nand[a, b]** shown with its two arguments, we have the evaluation of,

```
BooleanTable[Nand[a, b], {a, b}]
```

[1] For a "big Boolean Algebra" only the arguments T and F have to be examined.

returning **{False, True, True, True}** as the truth table. We see that these are the functional assignments to the four possibilities of the arguments to **Nand[ ]**. Use **Tuples[{True, False}, 2]** to see all four of these possibilities. The results from **BooleanTable[ ]** match up with the output from **Tuples[ ]**.

Or, alternatively we might write,

```
BooleanTable[BooleanFunction[7, 2]]
```

to extract the same truth table for **Nand[ ]**.

The first argument shown for **BooleanFunction[ ]** is the rule number for **Nand[ ]**, and the second argument tells us how many arguments are expected for the Boolean function. Or, if we are thinking in terms of a CA, the two arguments are the two colors for a cell, say, black and white.

A.3.2 How do we know that Nand is the 7th Boolean function?

How did we know that the first argument to **BooleanFunction[ ]** was a 7? In Volume I, Appendix A, Figure A.1, all 16 logic functions were laid out according to Wolfram's numbering scheme. The seventh one was **Nand[ ]** because the decimal number 7 is 0111 in binary representation.

These digits correspond to white, black, black, and black as the color for a cell as we rotate through in the order of BB, BW, WB, and WW as the two arguments to **Nand[ ]**. Thus, the little 2×2 tables in Figure A.1 were constructed according to the ordering returned by **Tuples[{B, W}, 2]**.

A.3.3 The easiest way for a CA to illustrate truth tables

In this section, I show the easiest way to exhibit the fact that a cellular automaton can output the truth table for the **Nand[ ]** logic function. Set up the initial conditions such that all four possibilities as arguments occur in sequence.

```
ArrayPlot[CellularAutomaton[{7, 2, 1/2},
                        {1, 1, 1, 0, 0, 1, 0, 0}, 1], Mesh → True]
```

The initial conditions are specified in the list **{1, 1, 1, 0, 0, 1, 0, 0}** indicating that the CA should start off as BB, BW, WB, and WW in blocks of two digits. This is set up to match the ordering from **Tuples[{B, W}, 2]**.

The list **{7, 2, 1/2}** specifies the rule number for **Nand[ ]**, $k = 2$ colors, black and white, while the neighborhood region is set at $r = 1/2$ so that the neighborhood range, the number of arguments to the logic function stays at $2r + 1 = 2$. Only one step in the evolution of the CA needs to be taken to determine the values in the truth table.

See the resulting diagram produced for this CA in Figure A.1 to verify that the updated cells at the next step for the initial conditions are, in fact, what we want. The colors of the second, fourth, sixth, and last cells at the first time step were generated as W, B, B, B, or 0111, or $FTTT$, as the truth table, the functional assignments to all four possibilities for the two arguments of **`Nand[ ]`**. Note that the neighborhood determining the color of the updated cell when **`1/2`** was specified is the cell above and the cell to its immediate left.

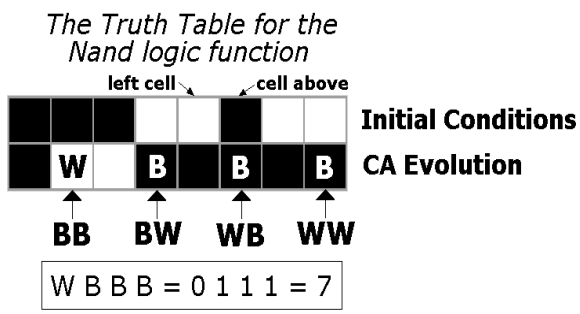

Figure A.1: *A CA reproducing the truth table for the* **`Nand[ ]`** *logic function.*

A.3.4 Verify with different logic function

Let's verify the result just found when the **`Nand[ ]`** logic function was taken as the specific example. Refer back to Figure A.1 of Appendix A, Volume I, and focus on the **`Xnor[ ]`** logic function. This occurs as the 9th Boolean function in Wolfram's numbering scheme, and therefore has the binary representation of 1001_2.

This list of digits **`{1, 0, 0, 1}`** is also the truth table for **`Xnor[ ]`** telling us what the functional assignment for all four possibilities will be, as verified through,

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

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

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

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

This was the notation used in Table 2.5, Chapter Two of Volume I, where the truth tables for all 16 two variable logic functions were presented. I also sometimes called the **`Xnor[ ]`** logic function `EQUAL` because it assigns T only when its two arguments are the same, otherwise it assigns F. In Figure A.1 of Volume I, $1001 \rightarrow TFFT$ is shown as black, white, white, black as we rotate through the four assignments BB, BW, WB, and WW.

Simply change the rule number from 7 to 9 in the above code. Everything else remains the same, there are only two colors, and the cell neighborhood is still defined as two cells, the one above and the one to the immediate left. Keep the initial conditions the same because, like before, we want to check what happens for BB, BW, WB, and WW.

Evaluating,

```
ArrayPlot[CellularAutomaton[{9, 2, 1/2},
                 {1, 1, 1, 0, 0, 1, 0, 0}, 1], Mesh → True]
```

will show that the colors of the updated cells at the first time step are black, white, white, and black when the relevant cells at the previous time step (the initial conditions) are BB, BW, WB, and WW. The expression for evolving a CA is also computing a truth table for the logic function specified as a Boolean function rule number. It's all very circular.

A.4 The Deconstruction (the harder part)

A.4.1 Select parts of the list produced by `CellularAutomaton`

The function **`CellularAutomaton[ ]`** by itself produces lists of 1s and 0s. For example, without an **`ArrayPlot[ ]`**,

```
CellularAutomaton[{7, 2, 1/2}, {1, 1, 1, 0, 0, 1, 0, 0}, 1]
```

returns,

```
{{1, 1, 1, 0, 0, 1, 0, 0}, {1, 0, 0, 1, 1, 1, 1, 1}}
```

The first list repeats the initial conditions, while the second list shows the evolution of the CA at the first time step.

Suppose we wanted to select the second element of the second list. Or, viewing the CA as an array, we want the cell in the second row and the second column. Use **`Part[ ]`**, the built-in function that extracts parts of a matrix,

```
Part[CellularAutomaton[{7, 2, 1/2},
                 {1, 1, 1, 0, 0, 1, 0, 0}, 1], 2, 2]
```

or, alternatively, through the syntactical short–cut,

```
CellularAutomaton[{7, 2, 1/2}, {1, 1, 1, 0, 0, 1, 0, 0}, 1][[2, 2]]
```

with both versions picking out the element 0, the first white cell from the left at the first time step.

A.4.2 Mapping a function across a list of arguments

Th next step in the deconstruction of the original opaque code is to rely on the ubiquitous **Map[ ]** command. In general, we would like to map some function across some number of arguments. First, though, we set up an anonymous pure function through **Function[ ]** that **Map[ ]** will use as its function that is to be mapped.

From our previous exposure to the syntax in **Function[***parameter, body***]**, we know that we first have to provide a *parameter* which will appear in the *body*. It is important to realize that the syntax demands that **Function[ ]** always have an *argument* appended at the end as in **Function[***parameter, body***][***argument***]** with the consequence that overall appearance of the function with its arguments is a bit surprising.

What will serve as the *body* for **Function[ ]**? It is that code just discussed in the last section that picks out the second element of the second list produced by **CellularAutomaton[ ]**.

What will serve as the *parameter* for **Function[ ]**? The one parameter will be the list of initial conditions that we have seen appearing as one of the arguments to **CellularAutomaton[ ]**.

What will be the argument for **Function[ ]**? The argument will be a list consisting of the colors of two previous cells as in **{1, 1}**.

Thus, we would write out,

```
Function[initial,
            (* the parameter *)
CellularAutomaton[{7, 2, 1/2}, initial, 1][[2, 2]]
            (* the body *)
                                                  ]
            (* close  Function *)
                                                    [{1, 1}]
            (* the argument to  Function *)
```

in order to return a 0, the second element of the second list when the initial condition for rule number 7 is **{1, 1}**. This tells us that a white cell was the update given that the two relevant cells were both black.

The more popular way of writing out a shorter version of a pure function is with the slot and ampersand notation. For the above, this version would look like,

```
CellularAutomaton[{7, 2, 1/2}, #, 1][[2, 2]]& [{1, 1}]
```

which must also return a 0. This result is merely the first entry for **Nand[ ]**'s truth table, $f_{12}(A = T, B = T) = F$, once again using the notation introduced in Chapter Two of Volume I, Table 2.5.

We want all four functional assignments for **Nand[ ]**. The argument list for the initial conditions of the CA running according to rule number 7 must be augmented to **{{1, 1}, {1, 0}, {0, 1}, {0, 0}}** in order to duplicate the TT, TF, FT, and FF arguments to the **Nand[ ]** logic function.

Now we are ready to introduce **Map [ ]** to accomplish this goal by mapping our derived anonymous function across the desired list of arguments,

```
Map[CellularAutomaton[{7, 2, 1/2}, #, 1][[2, 2]]&,
                                {{1, 1}, {1, 0}, {0, 1}, {0, 0}}]
```

As hoped for, this returns **{0, 1, 1, 1}**. All of the entries for Nand's truth table are now available. They confirm the entries in Table 2.5.

Now that the overall structure of the syntax is clear, we can add the short–cut for **Map[ ]** with,

```
CellularAutomaton[{7, 2, 1/2}, #, 1][[2, 2]]& /@
                            {{1, 1}, {1, 0}, {0, 1}, {0, 0}}
```

A.4.3 The use of offsets

There remains only one piece of the syntactic puzzle left to solve. This is the use of *offsets* to generalize the notion of the range of a cell's relevant neighborhood. When we have addressed this issue satisfactorily, we will have successfully cleared up the befuddlement over my first exposure to the tutorial on how to use CA to compute truth tables for logic functions.

Up to this point, we have specified $r = 1/2$ as the argument for the range of a cell's neighborhood. This means that the color of only two cells at the previous time step will influence the current cell scheduled to be updated. This must be the case since we are looking to find the truth table for a logic function with only two arguments.

The specification of $r = 1/2$ has appeared repeatedly in the argument list as **{7, 2, 1/2}**. According to the documentation, when $r = 1/2$, the *left* neighbor of the cell above together with the cell above define the relevant neighborhood.

We are going to replace and generalize the $r = 1/2$, a two cell neighborhood, or $r = 1$, a three cell neighborhood, or $r = 2$, a five cell neighborhood, (where in all these cases the neighboring cells are contiguous), with something called an *offset list*.

In our current example, replace $r = 1/2$ with an offset list of **{{0}, {1}}**. The offset **{0}** indicates that the cell above is the first relevant cell at the previous time step, and the offset **{1}** indicates that the cell *to the right* of the above cell instead of the cell *to the left* at the previous time step is the second relevant cell.

Therefore, the output of the CA evolving according to rule number 7 with two colors will change at the first time step for an arbitrarily specified initial condition list. In the previous examples, the two relevant cells were the cell above and the cell to its *left*, now with the new offset list, the two relevant cells are the cell above and the cell to the *right*.

The newly revised expression incorporating the offset list now looks like,

```
CellularAutomaton[{7, 2, {{0}, {1}}}, #, 1][[2, 2]]& /@
                                {{1, 1}, {1, 0}, {0, 1}, {0, 0}}
```

returning the list **{0, 1, 1, 1}**, which we see is the correct truth table for the Boolean function **Nand[ ]**.

One additional thing adding to the confusion is that, while the argument **1/2**, or the offset list **{{0}, {1}}**, indicating the cell neighborhood, does produce a different evolution for the CA, the resulting output of the truth table is the same. You just have to make sure that you are looking at the correct cell neighborhood for either specification. The rule associating the updated cell's color based on the two relevant cell's colors doesn't change.

Since we are on the topic, we might as well show some more examples of the offset list. The documentation shows an example using Rule 30 from Wolfram's 256 elementary cellular automata.

Since this is a logic function that has three arguments instead of two, we could simply rely on $r = 1$. This is the default specification for the number of relevant cells at the previous time step, so we wouldn't even explicitly have to include it, say, as **{30, 2, 1}**. But do so, and then set some arbitrary initial conditions,

```
CellularAutomaton[{30, 2, 1}, {1, 1, 0, 1, 0, 0}, 1]
```

Rule 30 running for one time step returns,

```
{{1, 1, 0, 1, 0, 0}, {1, 0, 0, 1, 1, 1}}
```

We can duplicate this default condition for the neighborhood cells with the offset list specifying three relevant cells, the cell above **{0}**, the cell to the immediate right **{1}**, and the cell to the immediate left **{- 1}**,

```
CellularAutomaton[{30, 2, {{- 1}, {0}, {1}}},
                                        {1, 1, 0, 1, 0, 0}, 1]
```

The capability of generalizing the cell neighborhood is illustrated next with the following change to the offset list.

```
CellularAutomaton[{30, 2, {{- 1}, {0}, {2}}},
                                        {1, 1, 0, 1, 0, 0}, 1]
```

The first two relevant cells are the same as before. However, the third relevant cell is not immediately to the right of the cell above, but skips one cell. The color

of all six cells at the first time step of the CA's evolution must then change. As before, this can be verified, by examining the new cell neighborhood at the previous time step, and realizing that the rule associating these colors, wherever they might be located, to the updated cell's color remains the same.

A.5 Adding My Own Obfuscations

Having gone to all this trouble of deciphering a less than transparent *Mathematica* expression, I decided to add a couple of extra layers of non–immediate apprehension. Remember, imagine that you are trying to make sense of these expressions after seeing them for the very first time.

A.5.1 Generalizing the arguments to the logic function

In my development above, the arguments to any two variable logic function were explicitly listed for use by **Map[ ]** as

```
{{1, 1}, {1, 0}, {0, 1}, {0, 0}}
```

It is easy however to replace this with **Tuples[{1, 0}, 2]**,

```
CellularAutomaton[{7, 2, {{0}, {1}}}, #, 1][[2, 2]]& /@
                                        Tuples[{1, 0}, 2]
```

A.5.2 Substituting any two variable logic function

We would immediately prefer to be able to compute the truth table for any of the sixteen two variable logic functions. Just replace the rule number argument **7** for **Nand[ ]** with the correct rule number for any other logic function. For example, suppose we were interested in inserting the rule number for **Implies[ ]**,

```
FromDigits[Boole[BooleanTable[Implies[a, b], {a, b}]], 2]
```

Continue to build up to this increasingly opaque expression,

```
CellularAutomaton[{FromDigits[Boole[BooleanTable[
  Implies[a, b], {a, b}]], 2], 2, {{0}, {1}}}, #, 1][[2, 2]]&
                                        /@ Tuples[{1, 0}, 2]
```

I'm sure it would be immediately obvious to you that, after glancing at this expression, it must return the correct truth table of **{1, 1, 0, 1}**.

A.5.3 Rampant circularity

We are not yet finished. We notice that it's not too hard to generalize, say, to finding the truth table for any logic function with four arguments. This would allow us to answer any question of the type: What is the functional assignment for $f_\star(A, B, C, D)$? Specifically what is the functional assignment for the $1,097^{th}$ logic function when the four variables assume the values of $A = T$, $B = F$, $C = T$, and $D = F$?

Modify the last expression for this situation as follows:

```
CellularAutomaton[{FromDigits[Boole[BooleanTable[
  BooleanFunction[1097, 4]]], 2, {{- 1}, {0}, {1}, {2}}}, #, 1]
                                    [[2, 2]]& /@ Tuples[{1, 0}, 4]
```

to find the functional assignment for $f_{1097}(A = T, B = F, C = T, D = F)$ from the resulting truth table.

To make our job of matching up the functional assignments with all sixteen variable settings easier, and so that we can pick out any setting we happen to be interested in, add a **Grid[ ]** command to the above expression,

```
Grid[Join[{Tuples[{T, F}, 4]},
                        above code for CellularAutomaton[ ]},
            ItemStyle → Directive[FontSize → 10, Blue],
                                Spacings →{1, 1}, Frame → All]
```

Finding **{T, F, T, F}** in the sixth column over from the left, we look down and see that its assignment according to the computed truth table is T. If we had happened to be interested in a different variable setting, say,

$$f_{1097}(A = T, B = F, C = F, D = T)$$

this is in the next column over in our grid, and the assignment here is F.

For my final effort, I will show you a function for displaying the functional assignments for a given Boolean function with arguments of the correct number of variables. It illustrates the typical way when using *Mathematica* you are led to more complicated looking code after first playing around with beginning cases.

```
functionAssignment[bfn_, nv_] :=
  Grid[Join[{Table[i, {i, 1, 2^nv]},
  {Tuples[{T, F}, nv]},
  {CellularAutomaton[{FromDigits[Boole[BooleanTable[
      BooleanFunction[bfn, nv]]], 2], 2,
      If[OddQ[nv], Table[{i, {i, - Floor[nv / 2, Floor[nv / 2]}],
      Table[{i}, {i, - ((nv /2) - 1), nv / 2]]}, #, 1][[2, 2]]&
  /@ Tuples[{1, 0}, nv] /. {0 → F, 1 → T}}], ItemSize → 7,
  ItemStyle → Directive[FontSize → 10, Blue], Frame → All]
```

To view all 16 function assignments for the above example, in other words the truth table for the logic function $f_{1097}(A, B, C, D)$, evaluate,

```
functionAssignment[1097, 4]
```

Look under column 6 labeled as T, F, T, F to see the functional assignment of T when the variables assume the values of $A = T$, $B = F$, $C = T$, and $D = F$.

What I was alluding to by "rampant circularity" is the fact that,

```
IntegerDigits[1097, 2, 16]
```

returns by itself, without the need for all the complicated computation by the CA in the above code, the correct digit list of sixteen 1s and 0s indicating the functional assignments.

But, in the end, just as it was in Wolfram's tutorial, the primary goal was to illustrate some of the finer syntactical points inherent in writing *Mathematica* code, especially when computing with CA. And, as mentioned before, it is nice to have an opportunity to address several key ideas involving logic functions, truth tables, decimal number expansions according to different bases, and cellular automata all wrapped up in one *Mathematica* expression.

Appendix B

Brown's Function Viewed As A Cellular Automaton

B.1 Introduction

I started off my introduction to Boolean Algebra in section 1.4, Volume I, with an example of a two variable Boolean function. This example was taken directly from Brown's similar example which was appropriately referenced at the beginning of the Supplemental Exercises for Chapter One. The only real difference between Brown and me was a notational one. These differences in notation were explored in the first five Supplemental Exercises.

But after the developments here in Chapter Sixteen and Appendix A, I think it might be worthwhile to spend a little time revisiting this Boolean function from the standpoint of a cellular automaton. Basically, it gives me another chance to review some of the interrelationships that exist between Boolean functions, truth tables, and cellular automata. It is not, in any sense, a "deep" analysis because all I do is concentrate mainly on translating Wolfram's numbering and coloring protocols for CA back to the Boolean functions.

It seems that Brown wanted to illustrate not only an example of a general functional assignment from a Boolean formula, but also a functional assignment from a "big Boolean Algebra." By this phrase, he meant defining a Boolean function from a carrier set with more elements than just T and F.

In Chapter One, Brown's function was shown to be a two variable Boolean function $f(x, y)$ with possible functional assignments coming from a carrier set defined, in my notation, as $\mathbf{B} = \{T, a, a', F\}$. Brown's original table of functional assignments as determined by the Boolean formula $a'x + ay'$ is reproduced at the top of the next page. This table is the same as a truth table as I have defined it, but Brown presents a different view of a truth table which is upcoming shortly.

B.2 Rearranging the Functional Assignment Table

Table B.1 below reproduces Brown's original functional assignment table using his notation, and proceeding down the sixteen rows with his chosen ordering. I have added row numbers, not in his original table, to make things clearer when referencing rows in the initial Supplementary Exercises of Chapter One.

Table B.1: *Brown's original functional assignment table. Here is his original caption for this table: "Function–table for $a'x + ay'$ over $\{0, 1, a', a\}$."*

Row	x	y	$f(x, y)$
1	0	0	a
2	0	1	0
3	0	a'	a
4	0	a	0
5	1	0	1
6	1	1	a'
7	1	a'	1
8	1	a	a'
9	a'	0	1
10	a'	1	a'
11	a'	a'	1
12	a'	a	a'
13	a	0	a
14	a	1	0
15	a	a'	a
16	a	a	0

Abstractly, of course, we are in the realm of a functional assignment from the template,

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

Instead of an expression looking like $f(a', 0) = 1$ as Brown would like to write it, and shown in row 9, we write instead $f(a', F) = T$.

We would furthermore prefer to rearrange the ordering of the assignment table in order to match up with the outcome from `Tuples[{T, a, a', F}, 2]`. Wolfram always shows black cells on the left progressing over to white cells on the right in a plot of any rule. So we want to follow this convention by an ordering progressing from T through F.

Thus, our rearranged table looks like Table B.2 at the top of the next page. We rotate through all sixteen variable assignments for x and y starting first with TT, and then Ta, and ending up with FF. The functional assignments $f(x, y)$ are the same as Brown's.

Table B.2: *My rearrangement of Brown's original table to correspond to the output from* **`Tuples[{T, a, a′, F}, 2]`** *and the different notation in the carrier set.*

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

B.3 An Appropriate CA

When we get around to using **`ArrayPlot[ ]`** for the one time step CA reproducing the truth table for this Boolean function, we will have to select *four* colors because we have a carrier set consisting of four elements. More importantly, the digits associated with a color will determine the rule number of the cellular automaton.

Let's use the following arbitrarily chosen color rules: (1) 3 is red is T, (2) 2 is green is a, (3) 1 is blue is a', and (4) 0 is yellow is F. These associations are laid out for reference in Figure B.1 at the top of the next page in an expansion of Table B.2. The creation of this table is an opportunity to practice how *Mathematica* constructs a two dimensional table with **`Grid[ ]`**.

Even though we have *four* elements in our carrier set, and *four* colors to keep track of in our eventual CA, we still only have a Boolean function depending on *two* arguments. The neighborhood of relevant cells must then consist of only two cells.

For example, two red cells occuring together in the initial conditions will produce a blue cell, $f(T,T) = a'$ from Row 1. A green cell to the left of a red cell will produce a yellow cell, $f(a,T) = F$ from Row 5, and so on.

Calculate the total number of possible functions as $4^{4^2} = 4,294,967,296$. We obtained this formula from the number of possible rules for a four color two neighbor cellular automaton, $k^{k^{(2r+1)}}$, where $k = 4$ and $r = 1/2$.

Row	x	y	Color x	Color y	$f(x, y)$	Color	Digits
1	T	T	R	R	a′	B	1
2	T	a	R	G	a′	B	1
3	T	a′	R	B	T	R	3
4	T	F	R	Y	T	R	3
5	a	T	G	R	F	Y	0
6	a	a	G	G	F	Y	0
7	a	a′	G	B	a	G	2
8	a	F	G	Y	a	G	2
9	a′	T	B	R	a′	B	1
10	a′	a	B	G	a′	B	1
11	a′	a′	B	B	T	R	3
12	a′	F	B	Y	T	R	3
13	F	T	Y	R	F	Y	0
14	F	a	Y	G	F	Y	0
15	F	a′	Y	B	a	G	2
16	F	F	Y	Y	a	G	2

Figure B.1: *An expansion of Table B.1 showing the colors used in the CA. This table was created through the* Mathematica **`Grid[ ]`** *function. Each actual color swatch is annotated with R, G, B, Y for the rendering in shades of gray.*

We know now that the correct rule number for our CA will be somewhere in the range of 0 to $4,294,967,295$. Furthermore, since there are four colors, the digits in the digit list corresponding to the correct rule will consist of 0, 1, 2, and 3 in any expansion of decimal number to base 4. The template for a rule number is then,

$$(d \times 4^{15}) + (d \times 4^{14}) + \cdots + (d \times 4^{1}) + (d \times 4^{0})$$

Rely on **`FromDigits[ ]`** to immediately find any decimal representation needed for a CA's rule number. For example, just enter an arbitrary list of sixteen digits,

```
FromDigits[{0, 3, 2, 1, 3, 0, 0, 2, 2, 0, 1, 1, 3, 2, 1, 1}, 4]
```

which returns a rule number 969 049 573. This particular rule number will dictate the evolution of some four color two variable CA.

But this list of digits submitted to **`FromDigits[ ]`** is very important. A selection of one of the four digits, together with its particular placement in the list, will determine the color of the functional assignment.

The digit 0 appearing first in the above arbitrarily chosen list matches up with the first element from **`Tuples[ ]`**, a TT, or a red cell above, and a red cell to its immediate left. This CA will produce a yellow cell because of the digit 0 when the cell above and the cell to its immediate left are both red, or $f(T, T) = F$. Thus, this arbitrary list cannot be the rule number we want for our Boolean function. When we provide the list of digits, we are literally providing the truth table of a Boolean function.

This example wants us to proceed backwards. We know the required functional assignment for each of the two arguments, and we have to fill in the correct digit d at the correct $(d \times 4^j)$ term in the base 4 expansion. The 0 as the first digit matching up with TT means that the first term in the expansion contributes (0×4^{15}) to the rule number. The second digit 3 will contribute (3×4^{14}) to the rule number, and so on.

Contrary to our arbitrarily filled in digit list, Brown's specific Boolean formula and resulting Boolean function will demand a different digit list. Two red cells must produce a blue cell. Thus, the first digit must be a 1 contributing (1×4^{15}) to our eventual rule number for the CA synonymous with Brown's Boolean function. A red cell and a green cell, that is the Ta appearing second in the **Tuples[ ]** ordering, must output another blue cell. The second digit is also a 1. We proceed in the same manner until we get to the final element from **Tuples[ ]**, an FF. Two yellow cells must produce a green cell. The final digit in the digit list is a 2, contributing a final $(2 \times 4^0) = 2$ to the rule number.

Working this out in full detail produces a list of sixteen digits, each of which must be a 0, 1, 2, or 3. Now use **FromDigits[ ]** to find the actual rule number for this Boolean function,

```
FromDigits[{1, 1, 3, 3, 0, 0, 2, 2, 1, 1, 3, 3, 0, 0, 2, 2}, 4]
```

indicating that rule number 1 594 515 210 is what we want. Check that the first two digits starting on the left are a 1 and 1, while the final digit is a 2.

We are now in possession of everything we need for **ArrayPlot[ ]** to visually confirm everything said above.

```
ArrayPlot[CellularAutomaton[
    {FromDigits[{1, 1, 3, 3, 0, 0, 2, 2, 1, 1, 3, 3, 0, 0, 2, 2}, 4],
                4, {{-1}, {0}}},
              {3, 3, 3, 2, 3, 1, 3, 0, 2, 3, 2, 2, 2, 1, 2, 0,
               1, 3, 1, 2, 1, 1, 1, 0, 0, 3, 0, 2, 0, 1, 0, 0}, 1],
    ColorRules → {0 → Yellow, 1 → Blue, 2 → Green, 3 → Red},
                                                    Mesh → True]
```

Let me provide you with a template so that you can orient yourself within this expression. **CellularAutomaton[ ]** has three arguments, as in,

CellularAutomaton[*rule number, initial conditions, steps***]**

But the first argument breaks down into three sub–arguments, contained in a list,

CellularAutomaton[{*rule number, number of colors, cell neighborhood* **},**
*initial conditions, steps***]**

The *rule number*, as just explained, is,

```
FromDigits[{1, 1, 3, 3, 0, 0, 2, 2, 1, 1, 3, 3, 0, 0, 2, 2}, 4]
```

The *number of colors* is **4**, and the cell neighborhood is a two cell neighborhood specified by the offsets, **{{- 1}, {0}}**, indicating that the cell above and the cell to its immediate left are to be consulted when a cell's color is to be updated. Pay close attention to the number and placement of the { and } brackets in all of this.

We are now ready to specify the *initial conditions* for this CA which will be in the form of some list indicating red, blue, green, and yellow cells. I chose the simple expedient of choosing the initial conditions by rotating through two red cells, followed by a red cell and a green cell, followed by a red cell and a blue cell, $\cdots$, and ending with two yellow cells.

```
{3, 3, 3, 2, 3, 1, 3, 0, 2, 3, 2, 2, 2, 1, 2, 0,
              1, 3, 1, 2, 1, 1, 1, 0, 0, 3, 0, 2, 0, 1, 0, 0}
```

This way I could easily check that the CA was producing the properly colored cell matching in order the elements from **Tuples[ ]**.

The final top–level argument for **CellularAutomaton[ ]** is to specify for how long the CA should run, that is, the number of *steps*. We require only one step to verify that the initial conditions are producing the properly colored cells at the next time step. In other words, all sixteen functional assignments $f(x, y)$ were checked as correct with the running of this CA with an appropriate rule number.

The remainder of the code are *options* to **ArrayPlot[ ]**. The most important option is the specification of the color rules for coloring any cell of the CA as explained above.

B.4 Constructing Tables with *Mathematica*

This section contains the code that created the expanded table for Brown's two variable Boolean function shown as Figure B.1. It consists mainly in manipulating lists to get them into the correct format that the **Grid[ ]** function demands. After that much has been accomplished, the rest of the code just specifies the appearance of the table.

Here is the full **Grid[ ]** expression before delving into the details.

```
Grid[Insert[Join[Table[{i}, {i, 16}], Tuples[{T, a, a′, F}, 2],
 Tuples[{Red, Blue, Green, Yellow}, 2], fa, colorfa, digitsfa, 2]
    {"Row", TraditionalForm[x], TraditionalForm[y], "Color x",
    {"Color y", TraditionalForm[f[x, y]], "Color", "Digits"}, 1],
    Spacings → {1, 1},
    ItemStyle → Directive[FontFamily → "Helvetica Neue",
                                          FontSize → 18],
    Dividers → {{True, True, False, True, False, True,
                                    False, False, True},
                          {True, {True, False, False, False}}},
    Background → {Lighter[Gray, .8], None},
    FrameStyle → Directive[Thickness[[3], Gray]]
```

Please notice, first of all, that this long expression is simply a filling out of the **Grid[ ]** function. The overall template looks like this,

Grid[{{···}, {···}, ··· , {···}}, *options***]**

The number of inner lists will appear as the number of rows in the table. The number of elements in each inner list will determine the number of columns in the table. Looking back at our table, we see that we want seventeen rows (counting the labeling for each column) and eight columns.

The constituent lists we want in the table are generated straightforwardly from **Table[ ]**, the two **Tuples[ ]**, and the separately declared global variables, the hand crafted lists **fa, colorfa, digitsfa**. They could have been defined within the **Grid[ ]** along with everything else, but we need to unclutter an already very long expression. Each **Tuples[ ]** generates two columns, so we have the right number of columns, that is, a total of eight columns.

The difficult part is when we **Join[ ]** these lists, the result in not in the correct format for **Grid[ ]** as exhibited in the above template. This is where the innocuous looking **2** as an option for **Join[ ]** is important. This directs the joining of the lists at level 2.

This is why **Table[ ], fa, colorfa, digitsfa** all had their individual entries as lists. For example, the hand crafted list of digits making up the rule number for the CA was created with,

```
digitsfa = {{1}, {1}, {3}, {3}, {0}, {0}, {2}, {2}, {1}, {1},
            {3}, {3}, {0}, {0}, {2}, {2}}
```

The net result after joining lists at level 2 was a first list of eight elements,

{{1, T, T, red, red, a′, blue, 1}

and a last list of eight elements,

{16, F, F, yellow, yellow, a, green, 2}}

The lists are now aligned the way **Grid[ ]** would like to see them. Another list for the column headings, explicitly shown in the above code, is inserted into this list at the first position with,

Insert[*just manipulated list***,** *column labels list***, 1]**

Now, all seventeen rows and eight columns of the table are appropriately arranged for,

Grid[{{···}, {···}, ··· , {···}}, *options***]**

Next follow all the *options* for the particular preferred appearance of the table. The only option which is always a bit tricky is **`Dividers`**. This specifies where any horizontal and vertical lines are desired in the table.

I think of the lines separating any columns as *vertical* lines, and likewise any lines separating the rows as *horizontal* lines. But this is the reverse of how the documentation describes them. We want five vertical lines explicitly drawn, two for the left and right borders with three more vertical lines separating the columns. But we also have to indicate where a vertical line should not appear. There are four places where a line should not be drawn.

All of this is accomplished with a list of nine **`True`** and **`False`**. Therefore, the list for the column specification is,

```
{True, True, False, True, False, True, False, False, True}
```

Turning now to the horizontal lines separating the rows, we have the top and bottom border, the line separating the column labels, and then a line separating every four variable assignments. But in this case it is not necessary to explicitly write out every single **`True`** and **`False`** specifying where a horizontal line should and should not appear. If there is a repetition, as there is for every four variable assignments, we need only write it out once as in **`{True, False, False, False}`**. The first **`True`** is necessary for the top border.

Thus, we have for the six horizontal lines drawn for our table, the list for the row specification, the list **`{True, {True, False, False, False}}`**.

Appendix C

Discrete Probability Distributions

C.1 Introduction

The hypergeometric discrete probability distribution has been discussed in some detail in the supplemental exercises to Chapter Six and Chapter Fifteen. This arose naturally because of our deep concern with how both Jeffreys and Jaynes brought it up when trying to deal with any inferential scenario that might conceivably be abstracted conceptually as sampling without replacement from an urn.

I was especially interested in exploring the consequences of having any assigned probabilities change with each trial. I wondered how to fit this concept into the framework developed throughout the initial introduction of the **Product Rule** that emphasized the *unchanging* nature of a probability at each trial when conditioned on the information under some model.

We relied on this concept to simplify,

$$P(A_t = a_i \,|\, A_{t-1}, A_{t-2}, \cdots, A_1, \mathcal{M}_k) \text{ to } P(A_t = a_i \,|\, \mathcal{M}_k)$$

In the case of sampling without replacement, the probabilities would seem to be decreasing at each trial. On the other hand, for the *Pólya Urn Scheme* the probabilities would seem to be increasing at each trial. Pondering this inferential situation, followed by reproducing Feller's answer for an example of the *Pólya Urn Scheme* as an application of De Finetti's Theorem, led me to a veritable thicket of discrete probability distributions.

After delving into the *Mathematica* documentation in order to understand how to apply **`HypergeometricDistribution[]`**, numerous side journeys presented themselves. This Appendix is an introductory presentation of some of these discrete probability distributions as seen by *Mathematica*.

C.2 Hypergeometric Distribution

By way of review, and to start off this enterprise, let's take another look at the discrete hypergeometric distribution. Many exercises in section 6.5 were exclusively devoted to the details of the hypergeometric distribution as this distribution was developed by Jaynes in his explanation of sampling without replacement. Exercises 15.3.12 and 15.3.13 also highlighted use of the hypergeometric distribution in the context of our discussion of how Jeffreys introduced it when he brought up the same topic of sampling without replacement discussed in Volume III.

`HypergeometricDistribution[]` has three arguments which, following the *Mathematica* documentation, we labeled as n, $nsucc$, and $ntot$. In the context of the urn scenario, $ntot$ was the total number of red and white balls in the urn, $nsucc$ was the total number of red balls in the urn, and n was the number of balls drawn from the urn.

So, if the inferential problem involved drawing five balls from an urn containing seven red balls and three white balls, the three arguments to the hypergeometric distribution over 0 and the positive integers would be,

$$n = 5$$

$$nsucc = 7$$

$$ntot = 10$$

However, **`HypergeometricDistribution[5, 7, 10]`** does nothing by itself. The *Mathematica* built-in symbol **`PDF[ ]`** standing for *probability density function* must be wrapped around any distribution in question. In *Mathematica* **`PDF[ ]`** is used for all distributions no matter whether they are density functions over some continuous domain or mass functions over some integer range. A second argument may be given to **`PDF[ ]`** to indicate the value where it is to be evaluated.

For our situation here with the hypergeometric distribution, we will label this argument as k. If the IP wanted the probability for drawing $k = 2$ red balls and $n - k$=3 white balls for a total draw of $n = 5$ from an urn consisting of $ntot = 10$ balls, $nsucc = 7$ of which are red, and $ntot - nsucc = 3$ of which are white, then evaluating,

```
PDF[HypergeometricDistribution[5, 7, 10], 2]
```

will return the probability of 1/12 for drawing two red balls and all three white balls.

If we want to see the actual formula that *Mathematica* is using to compute this probability, then substitute symbolic values into the above code,

```
PDF[HypergeometricDistribution[n, nsucc, ntot], k]
```

which evaluates to a combinatorial expression involving three **Binomial[ ]** terms,

```
(Binomial[nsucc, k] * Binomial[ntot - nsucc, n - k]) /
                                        Binomial[ntot, n]
```

I have taken the liberty of putting the arguments to **Binomial[ ]** not the way *Mathematica* orders them, but in the preferred way we would like to see them. This result is very satisfactory because it confirms both Jaynes's and Jeffreys's combinatorial formulas for the hypergeometric distribution.

Substituting the numerical values from the example,

$$
\begin{aligned}
P(k=2) &= \frac{\binom{7}{2} \times \binom{10-7}{5-2}}{\binom{10}{5}} \\
&= \frac{\frac{7!}{2!\,5!} \times \frac{3!}{3!\,0!}}{\frac{10!}{5!\,5!}} \\
&= \frac{21}{252} \\
&= \frac{1}{12}
\end{aligned}
$$

In order to obtain the entire distribution for the probabilities of drawing zero through five red balls, place the above code into a **Table[ ]** where the index k iterates from 0 through 5,

```
Table[PDF[HypergeometricDistribution[n, nsucc, ntot], k],
                                                  {k, 0, 5}]
```

which returns the list of six elements **{0, 0, $\frac{1}{12}$, $\frac{5}{12}$, $\frac{5}{12}$, $\frac{1}{12}$}**.

We see immediately that it is impossible to draw no or one red ball because then five or four white balls would have to be drawn and there are only three white balls in the urn. The two lowest probability events are where we would be lucky enough to draw out either all three white balls or all five red balls. You might want to revisit the details in Supplemental Exercise 6.5.6 at this point.

C.3 Beta Binomial Distribution

I discovered that the *Mathematica* documentation has a **GUIDE** section called *Urn Model Distributions* which seems to be right up our alley and worthy of further investigation. A distribution **BetaBinomialDistribution[ ]** is listed among many others. The brief description attached says: "number of white balls sampled from Pólya's urn"

Turning to the actual documentation for this discrete probability distribution, we find some interesting comments in the **Background & Context** section.

> The beta binomial distribution can be thought of as an abstraction of the Bernoulli and binomial distributions in which the success probability p of a known number of Bernoulli trials is random, and the associated binomial distribution has success probability p which follows the beta distribution. In Bayesian terms, this means that the beta binomial distribution arises as a posterior predictive distribution of a binomial variable in which the prior distribution on the success probability p is a beta distribution.

Just so, and it is nice to see the nod to the Bayesian interpretation when so little else of a Bayesian character is present in the *Mathematica* documentation. Nonetheless, a few comments are in order.

It is very important in the language and notation that we employ to maintain a strict distinction between a probability as written in the documentation with a p, and a numerical assignment under the information from some model $\mathcal{M}_k$. The "success probability" is, of course, the latter, and that is why I use a notation of $Q_i = P(X = x_i \,|\, \mathcal{M}_k)$ instead of a p or a θ which leads some people to think of the "success probability" as a "parameter." The notation q_i is used whenever an integration over all possible assignments to Q_i is to be performed as will eventually happen here for the *beta–binomial distribution*.

The assignment to Q_i results from the information implied by the *statement* $\mathcal{M}_k$. A p–like symbol is used only on the left hand side for an expression like $P(M_1, M_2)$, the computed prior predictive probability for future frequency counts. This is NOT the same as a p appearing in the binomial distribution on the right hand side.

I don't really like the language that says that p is "random." The probability density function for q, the *beta distribution*, must appear on the right hand side as part of the formal manipulation rules. It captures the IP's state of knowledge about q. It is through the mathematical procedure of multiplying the *binomial distribution* by the *beta distribution* followed by the integration over 0 to 1 that we arrive at the probability on the left hand side.

I hope I am not boring the reader by spelling this out for the umpteenth time. For the *beta–binomial distribution*, or what I prefer to call the prior predictive probability, we may write two probability expressions,

$$P(M_1, M_2) = \int_0^1 W(M)\, q^{M_1}\, (1-q)^{M_2}\, C_{\text{Beta}}\, q^{\alpha-1}\, (1-q)^{\beta-1}\, \mathrm{d}q$$

$$P\{S_n = k\} = \binom{n}{k} \int_0^1 \theta^k (1-\theta)^{n-k}\, F\{\mathrm{d}\theta\}$$

The first expression is my preferred notation for the prior predictive probability of future frequency counts, and the second is Feller's notation for de Finetti's Theorem.

Either of these expressions could be implemented in *Mathematica* with the built–in function for the *beta–binomial* discrete probability distribution that requires three arguments,

BetaBinomialDistribution[*alpha, beta, n***]**

The first two arguments are where we specify the two parameters for the *beta distribution*. The third argument is the number of draws from the urn. Let's start by putting this function through its paces for probabilities we know full well by now.

Suppose we want the probabilities of drawing k equal to zero through three red balls from the urn in $n = 3$ draws. If the IP is totally uninformed about every single aspect of the physical nature of the urns, the number of balls, and manner of drawing, then this state of ignorance is captured in the prior probability over model space by setting $\alpha = \beta = 1$. Evaluating,

```
Table[PDF[BetaBinomialDistribution[1, 1, 3], k], {k, 0, 3}]
```

returns the list of prior predictive probabilities $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$.

At the other extreme, if the prior probability over model space reflects a great deal of knowledge about the whole process of drawing balls from the urn, and suppose that this is reflected in a Dirac δ function assignment of $\delta(q - 1/2)$, then the prior predictive probabilities will follow the binomial distribution for future frequency counts with $\theta = 1/2$ (Feller's notation).

A plot of the probabilities for $k = 0$ through $n = 100$ draws from the urn will exhibit a symmetrical unimodal discrete version of a Gaussian curve with the peak at a probability of $k = 50$ red balls and $n - k = 50$ white balls. The Dirac δ function assignment of $\delta(q - 1/2)$ is approximated by setting, say, $\alpha = \beta = 1000$,

```
DiscretePlot[PDF[BetaBinomialDistribution[1000, 1000, 100], k],
   {k, 0, 100}, PlotRange → {{0, 100}, {0, .08}},
   ExtentSize → Full,
   AxesLabel → {Style["Number of red balls", 14], " "},
   ImageSize → Large]
```

When plotted at the top of the next page in Figure C.1, the approach to the binomial distribution and Normality can be easily discerned. For example, the plot exhibits a peak probability at $k = 50$ of 0.077. The probability on either side at $k = 49$ and $k = 51$ is 0.076.

Intermediate between these two states of knowledge would be a prior probability over the q assignments with, say, parameters $\alpha = \beta = 2$. A sample size of $n = 4$,

```
Table[PDF[BetaBinomialDistribution[2, 2, 4], k], {k, 0, 4}]
```

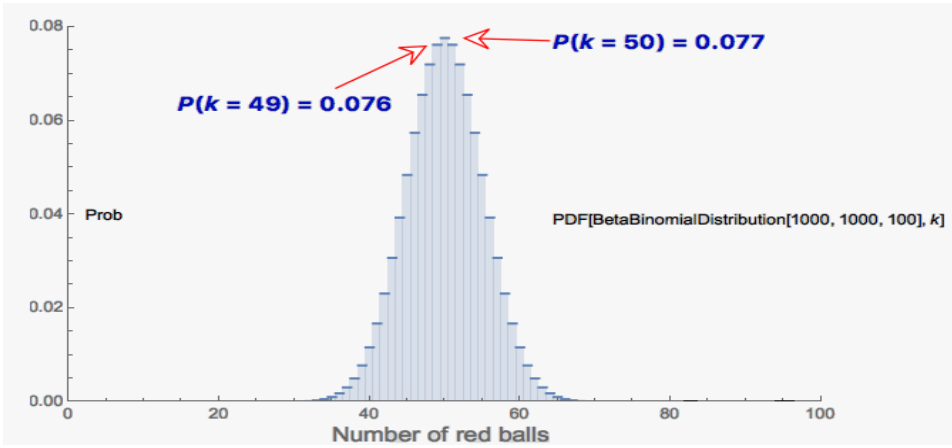

Figure C.1: *Plot of the beta binomial distribution showing the approach to both a binomial distribution with $\delta(q - 1/2)$ and a Normal distribution with $\mu = 50$.*

returns the list of prior predictive probabilities for k red balls and $n - k$ white balls that looks like this as k iterates through the values from 0 through 4,

$$\left\{\frac{1}{7}, \frac{8}{35}, \frac{9}{35}, \frac{8}{35}, \frac{1}{7}\right\}$$

Please take note that this changed prior probability for a numerical assignment to q by setting the parameters in the *beta distribution* to $\alpha = \beta = 2$ moved away from complete ignorance on the IP's part. It moved to a new state of knowledge that ever so slightly favored equal probability for drawing a red or white ball.

Nonetheless, this prior probability is still far removed from the Dirac δ function assignment of $\delta(q - 1/2)$. If the α and β parameters had remained at 1 for complete ignorance, then the list of five probabilities would all have been 7/35. But under the information from this new prior probability with $\alpha = \beta = 2$, the five probabilities are nudged slightly away from this uniform distribution in the expected direction.

It is very satisfying to verify this last result with a direct instantiation of the integration indicated in either of the two formulas given above. To that end, create this user defined function,

```
bbd[alpha_, beta_, n_] :=
   Table[Integrate[Binomial[n, k] θ^k (1 - θ)^(n - k),
   PDF[BetaDistribution[alpha, beta], θ], {θ, 0, 1}], {k, 0, n}]
```

Evaluating **bbd[2, 2, 4]** confirms the above result by returning the same list of probabilities.

What then does the symbolic formula as it is returned by *Mathematica* for the *beta–binomial distribution* look like? A bit of a surprise is in the offing because,

```
PDF[BetaBinomialDistribution[alpha, beta, n], k]
```

returns an expression with a strange symbol,

```
(Binomial[n, k]* Pochhammer[alpha, k] * Pochhammer[beta, n - k])
                                    / Pochhammer[alpha + beta, n]
```

I confess that I had never heard or come across a "Pochhammer symbol." But a quick check of the *Mathematica* documentation revealed it as something not so mysterious after all. **`Pochhammer[a, n]`** is defined as,

$$a \times (a+1) \times (a+2) \times \cdots \times (a+n-1)$$

For example, **`Pochhammer[10, 3]`** evaluates to,

$$10 \times 11 \times 12 = 1320$$

Exercise this new found mathematical symbol by computing the fourth prior predictive probability for the case of prior knowledge reflected by $\alpha = \beta = 2$. In the list given above this probability was 8/35. This is the probability for drawing $k = 3$ red balls and $n - k = 1$ white ball in a sample of four balls. Substituting the values for the binomial and the Pochhammer symbols results in,

$$\begin{aligned} P(k=3) &= \frac{\binom{4}{3} \times (2 \times 3 \times 4) \times 2}{4 \times 5 \times 6 \times 7} \\ &= \frac{4 \times 2}{5 \times 7} \\ &= 8/35 \end{aligned}$$

Another equivalent definition of **`Pochhammer[a, n]`** is,

$$\frac{\Gamma(a+n)}{\Gamma(a)}$$

with, for example, the above denominator **`Pochhammer[4, 4]`** since $\alpha = \beta = 2$ and $n = 4$ equal to,

$$\frac{\Gamma(a+n)}{\Gamma(a)} = \frac{\Gamma(8)}{\Gamma(4)} = \frac{7!}{3!} = 7 \times 6 \times 5 \times 4$$

From the definition of the *beta–binomial distribution*, it is obvious that this must be a special case of the prior predictive probability for future frequency counts. The special case must be where the dimension of the state space is $n = 2$. (Please note the confusion between the different notation for n.)

I wrote my own function **`priorPredictive[ ]`** to handle the *beta–binomial* situation for general n. And therefore it involved the Dirichlet distribution with n α_i parameters. The probabilities produced by my **`priorPredictive[ ]`** must be the same as the ones produced by **`BetaBinomialDistribution[ ]`**.

Evaluating,

```
Table[priorPredictive[{k, 4 - k}, {2, 2}], {k, 0, 4}]
```

does exactly that, confirming the equivalencies in the computation carried out by the two different functions.

Going back to the derivation that resulted in the formula incorporated into **priorPredictive[ ]**, isolate these terms to see the computational equivalency for, say, $M_1 = 3$ red balls and $M_2 = 1$ white ball in $M_1 + M_2 = M = 4$ draws from the urn, and $\mathcal{A} = \sum_{i=1}^{2} \alpha_i = 2 + 2 = 4$,

$$\frac{M!\,\Gamma(\mathcal{A})}{\Gamma(M+\mathcal{A})} = \frac{4!\,3!}{7!}$$

$$\frac{\Gamma(3+2)\times\Gamma(1+2)}{3!\,1!\,\Gamma(2)\,\Gamma(2)} = \frac{4!\,2!}{3!\,1!\,1!\,1!}$$

$$\begin{aligned} P(M_1 = 3, M_2 = 1) &= \frac{4!\,3!}{7!} \times \frac{4!\,2!}{3!\,1!\,1!\,1!} \\ &= \frac{4!\,2!}{7\times 6\times 5} \\ &= \frac{8}{35} \end{aligned}$$

We have managed to confirm numerically that the prior predictive probability for drawing three red balls and one white ball in four draws from the urn as was calculated by **priorPredictive[ ]** is the same probability as was calculated through **BetaBinomialDistribution[ ]**.

C.4 Pólya–Eggenberger Distribution

An example is presented under the **Applications** section of the Wolfram Language documentation for **BetaBinomialDistribution[ ]** which is called the "Pólya–Eggenberg urn distribution." Someone seems to have made a typo here on Pólya's collaborator, the name is "Eggenberger." More importantly, this discrete distribution as described in the documentation is exactly the same as studied in the supplemental exercises in section 15.3 as the *Pólya Urn Scheme*.

My version has arguments with the notation of b black balls, r red balls, c balls of the same color drawn, in a sample size of n.[1] I use this notation of n for the

[1] My apologies for making you switch the color of the balls in the urn from the previous examples using red and white balls.

sample size simply to correspond to Feller's and *Mathematica*'s usage. Ordinarily, in all other situations, n would refer to the dimension of the state space, where here the state space is, in my notation, $n = 2$.

Mathematica relies on Feller's definition of the *Pólya Urn Scheme* as I did as well in section 15.3 to create,

```
PolyaEggenbergerDistribution[b_, r_, c_, n_] :=
                        BetaBinomialDistribution[b / c, r / c, n]
```

Examine the symbolic expression for the distribution when symbolic arguments are inserted, and after evaluating,

```
PDF[PolyaEggenbergerDistribution[b, r, c, n], k]
```

The expression returned for the definition of the probability distribution consists of the same kind of mixture of **Binomial[ ]** and **Pochhammer[ ]** symbols that we have seen before,

```
(Binomial[n, k] * Pochhammer[b / c, k] * Pochhammer[r / c, n - k])
                                    / Pochhammer[b / c + r / c, n]
```

for $0 \leq k \leq n$.

Before examining a numerical example using these expressions, compute all of the prior predictive probabilities from first principles. Refer back to Supplemental Exercise 15.3.9 and Table 15.4 to refresh your memory as to how these calculations were organized.

As in these previous examples, use this notation. The number of black balls in the urn is $b = 7$, the number of red balls is $r = 3$, the number of balls replaced of the same color as drawn is $c = 2$, and the sample size is $n = 3$. The example there studied $c = 1$ where *one* ball of the same color drawn was added to the urn, while this example takes $c = 2$ where *two* balls of the same color drawn are now added to the urn. We require the probabilities for $k = 0$ through $k = 3$ black balls to be drawn from the urn.

Evaluating,

```
Table[N[PDF[PolyaEggenbergerDistribution[7, 3, 2, 3], k]],
                                                    {k, 0, 3}]
```

returns this list of four prior predictive probabilities,

```
{0.0625, 0.1875, 0.3375, 0.4125}
```

for (1) all three red balls, (2) one black ball and two red balls, (3) two black balls and one red ball, and (4) all three black balls.

Check the veracity of these four probabilities as returned by our definition of the **`PolyaEggenbergerDistribution[ ]`**. As we said, we will accomplish this goal by adhering to the actual prescription given by Feller.

It turns out that the numerator for any specific sequence for any given k will always possess the same three terms, while the denominator in all cases will always possess the three terms $10 \times 12 \times 14$. The numerator will also have a binomial factor reflecting the number of different ways in which the sequence of black and red balls could have been drawn.

For $k = 0$ and $n - k = 3$, the first prior predictive probability in the list, we have the case of three red balls and no black balls. This probability, from first principles, is simply calculated, with $c = 2$ red balls being added to the mix every time a red is drawn, as

$$P(rrr) = 1 \times \frac{3}{10} \times \frac{5}{12} \times \frac{7}{14}$$

$$= 0.0625$$

The first term is the binomial factor $\binom{n}{k} \equiv \binom{3}{0} = 1$ for just the one way that three red balls could be drawn from the urn.

For $k = 1$ and $n - k = 2$, the second prior predictive probability in the list, we have the case of two red balls and one black ball. This probability, from first principles, is just as easily calculated as,

$$P(1b, 2r) = 3 \times \frac{3}{10} \times \frac{5}{12} \times \frac{7}{14}$$

$$= 0.1875$$

The first term is the binomial factor $\binom{n}{k} \equiv \binom{3}{1} = 3$ for the three ways that one black ball and two red balls could be drawn from the urn.

For completeness, let's not skip over any of the details in this second probability calculation. The numerator was written as $3 \times 5 \times 7$, but this particular order would make sense only for the particular sequence of $P(rrb) = \frac{3}{10} \times \frac{5}{12} \times \frac{7}{14}$. A second sequence is $P(rbr) = \frac{3}{10} \times \frac{7}{12} \times \frac{5}{14}$, with the third and final sequence, $P(brr) = \frac{7}{10} \times \frac{3}{12} \times \frac{5}{14}$. The terms in the denominator are the unchanging $10 \times 12 \times 14$ no matter the sequence. So we see that the computation carried out above is correct,

$$P(1b, 2r) = 3 \times \left(\frac{3}{10} \times \frac{5}{12} \times \frac{7}{14} \right)$$

$$= 0.1875$$

Apply the same detailed analysis to the next prior predictive probability in the

list, the probability for two black balls and one red ball, $k = 2$ and $n - k = 1$.

$$P(2b, 1r) = 3 \times \frac{3}{10} \times \frac{7}{12} \times \frac{9}{14}$$

$$= 0.3375$$

The first term is the same binomial factor $\binom{n}{k} \equiv \binom{3}{2} = 3$ for the three ways that two black balls and one red ball could be drawn from the urn.

The sequence bbr has the numerator $7 \times 9 \times 3$, the sequence brb has the numerator $7 \times 3 \times 9$, while the third sequence rbb has the numerator $3 \times 7 \times 9$. The denominator for all three possible sequences is still $10 \times 12 \times 14$.

To finish up, the final prior predictive probability in the list is the probability for $k = 3$ black balls and $n - k = 0$ red balls. There is only one possible way for this sequence to occur, so the multiplicity factor represented by $\binom{3}{3} = 1$.

$$P(bbb) = 1 \times \frac{7}{10} \times \frac{9}{12} \times \frac{11}{14}$$

$$= 0.4125$$

Now that we possess a thorough familiarity with all of the numbers involved in calculating these probabilities, investigate the more mysterious Pochhammer symbols in the definition of **`PolyaEggenbergerDistribution[ ]`**. Focus on the case of $k = 1$ black ball and $n - k = 2$ red balls for which we know the probability is 0.1875.

In the numerator we will have,

```
Binomial[n, k] * Pochhammer[b / c, k] * Pochhammer[r / c, n - k]
```

and after substituting,

`Binomial[n, k]` $= \binom{3}{2} = 3$

`Pochhammer[b / c, k]` $= \frac{7}{2}$

`Pochhammer[r / c, n - k]` $= \frac{3}{2} \times \frac{5}{2} = \frac{15}{4}$

In the denominator, we will have,

```
Pochhammer[b / c + r / c, n]
```

and after substituting,

`Pochhammer[b / c + r / c, n]` $= 5 \times 6 \times 7 = 210$

Putting everything back together again,

$$P(1b, 2r) = \frac{3 \times (\frac{7}{2} \times \frac{15}{4})}{210} = 0.1875$$

The same numbers that appeared in the course of the above analysis can be reproduced from first principles. Keep the individual numerical values for the Pochhammer symbols, move the 2s underneath, and then distribute the three 2s appropriately in front of the 5, 6, and 7 so that we have,

$$\begin{aligned} P(1b, 2r) &= \frac{3 \times (7 \times 3 \times 5)}{2 \times 2 \times 2 \times 5 \times 6 \times 7} \\ &= \frac{3 \times (3 \times 5 \times 7)}{(2 \times 5) \times (2 \times 6) \times (2 \times 7)} \\ &= 3 \times \frac{3 \times 5 \times 7}{10 \times 12 \times 14} \\ &= 3 \times \left(\frac{3}{10} \times \frac{5}{12} \times \frac{7}{14} \right) \end{aligned}$$

To repeat, even though this ordering shows the probability for rrb, the computation of the probability for the other two possibilities, brr and rbr, is exactly the same because,

$$\left(\frac{3}{10} \times \frac{5}{12} \times \frac{7}{14} \right) = \left(\frac{7}{10} \times \frac{3}{12} \times \frac{5}{14} \right) = \left(\frac{3}{10} \times \frac{7}{12} \times \frac{5}{14} \right)$$

C.5 Pólya Distribution

For the following discussion, we will rely on Feller's explanation and formula for the *Pólya distribution* [6, pg. 121, pg. 142, Volume I]. Let's verify numerically the first part of Feller's Equation (2.4),

$$p_{n_1,n} = \frac{\binom{n_1 - 1 + b/c}{n_1} \binom{n_2 - 1 + r/c}{n_2}}{\binom{n - 1 + (b+r)/c}{n}}$$

Make the translation over to our simplified notation of $P(2b, 1r)$ for the probability of drawing two black balls and one red ball with $n = 3$, $n_1 = 2$ and $n_2 = 1$. Of the total of ten balls in the run, there are $b = 7$ black balls and $r = 3$ red balls with $c = 2$ balls of the same color drawn being replaced,

$$P(2b, 1r) = p_{2,3} = \frac{\binom{2-1+7/2}{2}\binom{1-1+3/2}{1}}{\binom{3-1+10/2}{3}}$$

$$= \frac{\binom{9/2}{2}\binom{3/2}{1}}{\binom{7}{3}}$$

$$= 0.3375$$

I went directly to **Binomial[ ]** for the result,

```
N[(Binomial[9/2, 2] * Binomial[3/2, 1]) / Binomial[7, 3]]
```

In even more detail, we might wonder what happens with the inevitable $\sqrt{\pi}$ terms that must appear in factorials involving 1/2. For example, in the first combinatorial expression, the $\sqrt{\pi}$ appearing in both the numerator and the denominator cancel,

$$\binom{9/2}{2} = \frac{(9/2)!}{(5/2)!\,2!} = \frac{\frac{945\sqrt{\pi}}{32}}{\frac{15\sqrt{\pi}\times 2}{8}} = \frac{63}{8}$$

The second version of the combinatorial formula that Feller presents right after this one is,

$$p_{n_1,n} = \frac{\binom{-b/c}{n_1}\binom{-r/c}{n_2}}{\binom{-(b+r)/c}{n}}$$

Verify that the same probability for two black balls and one red ball is computed through,

$$P(2b, 1r) = p_{2,3} = \frac{\binom{-7/2}{2}\binom{-3/2}{1}}{\binom{-10/2}{3}}$$

$$= 0.3375$$

Rely once again on a straightforward implementation of **Binomial[ ]**,

```
N[(Binomial[-7/2, 2] * Binomial[-3/2, 1]) / Binomial[-10/2, 3]]
```

The *Mathematica* version of the *Pólya distribution* is slightly more opaque, and I doubt that anyone coming to it afresh without the extensive groundwork we have undertaken here could make hide nor hair of it. Here is the symbolic expression returned by **PDF[PolyaDistribution[p, α, n] , k]** similar to previous ones we have looked at,

```
(Binomial[n, k] * Pochhammer[(1 - p) / α, n - k] *
        Pochhammer[p / α, k]) / Pochhammer[(1-p) / α + p / α, n]
```

Curiously, *Mathematica* doesn't automatically simplify the first argument to the **Pochhammer** symbol in the denominator, so we have to apply a **Simplify[ ]** to obtain,

```
(Binomial[n, k] * Pochhammer[(1 - p) / α, n - k] *
            Pochhammer[p / α, k]) / Pochhammer[1 / α, n]
```

Mathematica defined the *Pólya distribution* as a *beta–binomial distribution* by transforming its first two arguments,

```
PolyaDistribution[p_, α_, n_] :=
                BetaBinomialDistribution[p / α, (1-p) / α, n]
```

So, now we have to figure out the relationship between the correct arguments to the **BetaBinomialDistribution[ ]** and the corresponding arguments to **PolyaDistribution[ ]**.

We know that **BetaBinomialDistribution[b / c, r / c, n]** returns the correct probabilities. Recall that two balls of the same color as just drawn were replaced in the urn so that $c = 2$. Also, there are $b = 7$ black balls in the urn together with $r = 3$ red balls.

Thus, the three arguments for,

```
BetaBinomialDistribution[7 / 2, 3 / 2, 3]
```

would correspond to the three arguments in **PolyaDistribution[.7, .2, 3]**.

To confirm this,

```
Table[PDF[PolyaDistribution[.7, .2, .3], k], {k, 0, 3}]
```

does, in fact, return the correct probabilities in the list,

```
{0.0625, 0.1875, 0.3375, 0.4125}
```

for zero through three black balls in three draws from the urn consisting of seven black balls and three red balls.

Appendix D

Syntax of *Mathematica* Symbolic Expressions

D.1 Introduction

The basic idea of a general and abstract *Mathematica* symbolic expression was briefly introduced in Appendix A of Volume I with an example of the logic function **And[p, q]**.

There, I used the template below as the notation for *Mathematica*'s overall syntactic structure for symbolic expressions,

$$\texttt{head[}arg_1\texttt{, }arg_2\texttt{, }\cdots\texttt{, }arg_n\texttt{]}$$

where **And** was the **head** with **p** and **q** as the two arguments. I will now change this particular notation in light of the upcoming discussion to,

$$\texttt{h[}a_{h1}\texttt{, }a_{h2}\texttt{, }\cdots\texttt{, }a_{hn}\texttt{]}$$

D.2 Nested Syntax

My intent in this Appendix is to go into a little more detail of this syntactic structure for symbolic expressions. An essential component is the idea of allowable unlimited nesting within the expressions. This means that any argument a_{hi} appearing in,

$$\texttt{h[}a_{h1}\texttt{, }a_{h2}\texttt{, }\cdots\texttt{, }a_{hn}\texttt{]}$$

could itself be in the form of another head with its own arguments.

Suppose then that a_{h2}, corresponding to this idea of nesting, is now written as **f[**a_{f1}**,** a_{f2}**]** so that the full symbolic expression becomes,

$$\mathtt{h[}a_{h1}\mathtt{,\ f[}a_{f1}\mathtt{,}a_{f2}\mathtt{],\ \cdots,\ }a_{hn}\mathtt{]}$$

Taking the nesting concept one step further, suppose the second argument to **f[ ]**, namely a_{f2}, is **g[**a_{g1}**]**. The full expression, illustrating the beginning stages of a nested framework, looks like,

$$\mathtt{h[}a_{h1}\mathtt{,\ f[}a_{f1}\mathtt{,\ g[}a_{g1}\mathtt{]],\ \cdots,}a_{hn}\mathtt{]}$$

The expressions **h[ ], f[ ], g[ ]** might represent either *Mathematica* built–in functions, or user–defined functions.

These expressions are evaluated according to *Mathematica*'s specific technical implementation returning, say, some expression e_i. Suppose that **f[**a_{f1}**,** a_{f2}**]** is a user supplied function and **h[ ]** is some *Mathematica* built–in function. Then, **f[**a_{f1}**,** a_{f2}**]** might evaluate to e_2.

Mathematica keeps on evaluating the resulting overall expression for as long as it is able. Thus, if it can evaluate **h[**a_{h1}**,** e_2**, ⋯,** a_{hn}**]**, it will perform this evaluation next.

D.3 Examples

Here is a simple example of nesting as it naturally occurs in a *Mathematica* symbolic expression which I repeat from Wolfram's own early explanation.[1] The symbolic expression **1 + 2 x** written in **FullForm[ ]** is **Plus[1, Times[2, x]]**.

This is observed to follow the above template through,

$$\mathtt{h[}a_{h1}\mathtt{,\ }a_{h2}\mathtt{]} \equiv \mathtt{h[}a_{h1}\mathtt{,\ f[}a_{f1}\mathtt{,\ }a_{f2}\mathtt{]]}$$

where **h** is the *head* **Plus**, **f** is the *head* **Times**, both obviously built–in functions, with the first argument a_{h1} equal to 1, and the second argument a_{h2} equal to **Times[2, x]**. The two arguments a_{f1} and a_{f2} are obvious.

Give a specific value to the symbol **x** through,

$$\mathtt{1 + 2\ x\ /.\ x} \rightarrow \mathtt{3}$$

The evaluation order of this expression proceeds as follows: (1) x assigned value 3, (2) $1 + (2 \times 3)$, (3) $2 \times 3 = 6$, and the final output, (4) $1 + 6 = 7$.

[1] *S. Wolfram, The Mathematica Book, 4th Edition, Wolfram Research, Champaign, IL, 2004.*

This presents us with another opportunity to parse the **FullForm[ ]** symbolic expression to see if it follows the nesting template. Applying **FullForm[ ]** to the short–cut syntax, we find that the symbolic expression looks like this,

```
ReplaceAll[Plus[1, Times[2, x]], Rule[x, 3]]
```

At the top–most level, there is just the *head* **h**, **ReplaceAll**, and its two arguments a_{h1} and a_{h2}. The first argument a_{h1} is **Plus[1, Times[2, x]]**, while the second argument a_{h2} is **Rule[x, 3]**. Shifting our thought processes to the infinite nesting structure, the first argument a_{h1} is also **f[**a_{f1}**,** a_{f2}**]**. The *head* **f** is **Plus** with its own two arguments a_{f1} and a_{f2}.

Argument a_{h2} is **gstar[**a_{gstar1}**,** a_{gstar2}**]** with *head* **gstar** equal to **Rule** and its two arguments a_{gstar1} and a_{gstar2} equal to **x** and **3**. The head **g** was reserved for **Times** and its two corresponding arguments a_{g1} and a_{g2}.

D.4 Logic Functions

Returning to logic function symbolic expressions in *Mathematica*, consider the full DNF for forward implication $A \rightarrow B$,

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

The nesting template for all of *Mathematica*'s symbolic expressions is not difficult to discern in this case. We have at the top–most level,

h[a_{h1}**,** a_{h2}**,** a_{h3}**]**

with *head* **h** equal to **Or** and each of the three arguments equal to **f[**a_{f1}**,** a_{f2}**]** with *head* **f** equal to **And**. The arguments to **And[ ]** must be either **True** or **False**.

Replacing the symbols **p** and **q** with Boolean values through,

```
Or[And[p, q], And[Not[p], q], And[Not[p], Not[q]]] /.
                                    {p → True, q → False}
```

evaluates to **True** as verified by Table 2.5 in Volume I.

The template for general symbolic expressions in *Mathematica* is evident in the **FullForm[ ]** expansion of the above short–cut syntax,

```
ReplaceAll[Or[And[p, q], And[Not[p], q], And[Not[p], Not[q]]],
              List[Rule[p, True], Rule[q, False]]]
```

The nesting and evaluation procedures as used by *Mathematica* are essentially the same as shown in Chapter Two when I used my own functional notation of $f_j(A, B)$ to prove tautologies. One difference in my approach was that each logic function could, by definition, possess only two arguments.The example on pg. 47 of Volume I proved that the logic function `XOR` $A \oplus B$ was logically equivalent to the DNF expression $(A \wedge \overline{B}) \vee (\overline{A} \wedge B)$.

Work through a similar exercise in my notation proving the logical equivalency of the full DNF for the `NAND` logic function and $A \uparrow B$. I alleviate the tediousness of solving these exercises by hand with the eventual pleasure of watching the initially bloated expressions computationally collapse down into the correct solution.

First though, let's verify that *Mathematica* confirms the logically equivalency through an application of **`TautologyQ[ ]`**.

```
TautologyQ[Equivalent[Nand[p, q],
    Or[And[p, Not[q]], And[Not[p], q], And[Not[p], Not[q]]]]]
```

does return **`True`**.

I introduced the topic of logical equivalency back in section 2.6.1 of Chapter Two, Volume I. Abstractly, a tautology is the logical equivalency between two different logic expressions which I wrote as $\mathbf{F_{Old}} \leftrightarrow \mathbf{F_{New}}$. The question then becomes whether $(A \uparrow B) \leftrightarrow ((A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B}))$ evaluates to T for all four possible variable assignments to A and B.

Now for the promised hand evaluation of nested functions, we revert to my functional notation of Chapter Two for logic functions. For the first possible variable assignment of $A = T, B = T$,

Step 1 $(A \uparrow B) \leftrightarrow ((A \wedge \overline{B}) \vee (\overline{A} \wedge B) \vee (\overline{A} \wedge \overline{B}))$
Step 2 $f_8\,[\,(f_{12}(T,T), f_{15}\,[\,f_{15}[f_5(T, f_7(T,T)), f_5(f_6(T,T), T)\,], f_5(f_6(T,T), f_7(T,T))\,]\,]$
Step 3 $f_8\,[\,(F, f_{15}\,[\,f_{15}\,[\,f_5(T,F), f_5(F,T)\,], f_5(F,F)\,]\,]$
Step 4 $f_8\,[\,(F, f_{15}[f_{15}(F,F), F\,]\,]$
Step 5 $f_8\,[\,F, f_{15}(F,F)\,]$
Step 6 $f_8(F,F) \rightarrow T$

We would have to proceed through with the other three variable assignments showing that they also collapse down to T to prove that logical equivalency exists between the two forms.

D.5 Diversity of Appearance for a Head

A surprising generality associated with a *head* is that it is not necessarily restricted to a single symbol as I have been illustrating it up to this point with, say, **`h[ ]`**, **`Rule[ ]`**, or **`And[ ]`**. In Volumes II and III, I mentioned that the *head* for a pure anonymous function and its one *argument* returning the value y^2 looks like this,

```
Function[x, Power[x, 2]][y]
```

where now the entire *head* is **Function[x, Power[x, 2]]** and its one argument is **[y]**. **x** is called the *parameter* and **Power[x, 2]** is called the *body*.

This strange format for the *head* is somewhat hidden away in the most common application of a pure function within a **Map[ ]**. For example,

```
Map[Function[x, Power[x, 2]], List[a, b, c]]
```

evaluates to $\{a^2, b^2, c^2\}$. Each single argument to the anonymous function now resides in the list of arguments.

Most often, you will not see the fully explicit symbolic expressions as I have outlined then here, but rather in the form of syntactical short–cuts. For example, the above fully explicit symbolic expression involving **Map[ ]** would more likely appear as,

```
#^2& /@ {a, b, c}
```

Mathematica employs even stranger formats for *heads*. In addition to the *head*, *argument* appearance of **Function[ ][ ]**, there is also the fully explicit symbolic expression for derivatives which has a head like **Derivative[2][f]** together with its argument **[x]** for the second derivative $f''(x)$. Refer back to my more detailed discussion of symbolic expressions for derivatives like,

```
Derivative[1, 1][f][x, y]
```

in Appendices B and C of Volume III. These fully explicit symbolic expressions for derivatives were discussed within the context of Information Geometry and the Maximum Entropy Principle.

Finally, a *head* might assume an even stranger appearance like **(f + g)** with an argument **[x]** which, according to Wolfram, represents the addition of two operators within the *head*, so that **(h + j)[z]** is **h[z] + j[z]**.

Bibliography

[1] Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*, John Wiley & Sons, Ltd., Chichester, England, 1994.

[2] Blower, David J. *Information Processing: Boolean Algebra, Classical Logic, Cellular Automata, and Probability Manipulations. Volume I, First Revised Edition*, CreateSpace, Amazon.com, October 2017.

[3] Blower, David J. *Information Processing: The Maximum Entropy Principle. Volume II*, CreateSpace, Amazon.com, June 2013.

[4] Blower, David J. *Information Processing: An Introduction to Information Geometry. Volume III*, CreateSpace, Amazon.com, April 2016.

[5] Brown, Frank Markham. *Boolean Reasoning: The Logic of Boolean Equations*, Second Edition. Dover Publications, Mineola, NY, 2003.

[6] 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.

[7] 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.

[8] Garrett, Anthony J. M. Whence the Laws of Probability? in *Proceedings of the* 17th *International Workshop in Maximum Entropy and Bayesian Methods*, Boise, ID, 1997, pp. 71–86, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[9] Geisser, Seymour. *Predictive Inference: An Introduction*. Chapman & Hall, New York, NY, 1993.

[10] Hofstadter, Douglas R. *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books, New York, 1979.

[11] Jaynes, Edwin T. *Probability Theory: The Logic of Science*, ed. by G. Larry Bretthorst, Cambridge University Press, New York, NY, 2003.

[12] 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.

[13] Jeffreys, Harold. *Scientific Inference*, Second Edition, Cambridge University Press, 1957.

[14] Jeffreys, Harold. *Theory of Probability*, Third Edition, Oxford University Press, 1961.

[15] Mendelson, Elliott. *Boolean Algebra and Switching Circuits*, Schaum's Outline Series in Mathematics, McGraw–Hill, New York, NY, 1970.

[16] Stoll, Robert R. *Set Theory and Logic*, Dover Publications, Mineola, NY, 1979.

[17] Tipler, Frank J. *The Physics of Immortality: Modern Cosmology, God and the Resurrection of the Dead*, Doubleday, New York, NY, 1994.

[18] Wolfram, Stephen. *A New Kind of Science*, Wolfram Media, Inc., Champaign, IL, 2002.