DISCRETE PROBABILITY THEORY FOR UNDERGRADUATES

TABLE OF CONTENTS

1 Discrete Probability Spaces

In this chapter we introduce the mathematical objects encoding both the events observed in an experiment and the likelihood of their occurrences. Such objects are characterized by triples (Ω, β, P), consisting of a *countable set* Ω, the set of its *subsets* β, and a *set function* $P : \beta \to [0, 1]$ satisfying a list of three axioms. Accordingly, we have occasion here to begin our study of discrete probability theory by introducing the foundational aspects required to understand this object. The reader may object such generality is unnecessary, especially for the purposes of an undergraduate, however, such objections would be naive, for they neglect to consider the necessity of such generality in several both major and modern-day applications.

1.1 Set Theory

In this section, we introduce enough set theory to study the main ideas of this text.

Definition 1. *We say a collection of elements, denoted S, is a set. Furthermore, we say any not necessarily proper subcollection of elements, say T, is a subset. We denote the subset relation among sets by $\subset$ and write $T \subset S$. We write $|S|$ for the number of elements in S, and refer to this number as its cardinality.*

In terms of elements, the subset relation is satisfied by the condition that every element t of T, or in notation, $t \in T$, that we also have $t \in S$. The subset relation between sets allows us to define the equality of these objects.

Definition 2. *Let S and T be sets, then we say S is equal to T and write $S = T$ if and only if both $T \subset S$ and $S \subset T$ are true.*

Given this definition, one must demonstrate both conditions equivalent to the equality of two sets hold in order to demonstrate that they are equal. We shall practice this requirement in both examples and exercises.

Next, we shall introduce the implements at a mathematician's disposal for generating sets from those that they are given, in much the same way that one is able to generate sets of numbers from a subset of the same by means of arithmetic operations. Indeed, we shall introduce *set operations* for this purpose. They are named according to the arithmetic operations that they generalize.

Definition 3. *Let S and T be sets. The we say the set of elements in either S or T, denoted $S \cup T$, is their union. Furthermore, we say the set of elements in both S and T, denoted $S \cap T$, is their intersection.*

There is other notation common in the literature used to describe these operations. One reads

$$S \cup T = \{\, x \,|\, x \in S \;\text{ OR }\; x \in T\}$$

as elements x 'such that' for '$|$' in either S or T. This elemental notation easily generalizes to other sets and we shall use it throughout the text. Indeed, $S \cap T = \{\; x \,|\, x \in S \;\text{ AND }\; x \in T\}$. One might notice the curious capitalization of the conjunctions 'or' and 'and' that we used in the defining conditions aspect of our set notation. The reason for this is to emphasize that these conditions are defined logically. We shall review what we need of formal logic below and its relation to set theory, however, the relation may perhaps already be clear to the learned reader.

In terms of a comparison to arithmetic operations, these set operations we have defined correspond to the sum and product operations on the set of integers, and for that reason, are occasionally referred to as the *logical sum* and *logical product*, respectively. If one should wonder whether there are operations corresponding to the arithmetic inverses, namely differences and quotients, this is a rather more delicate question, especially in the second case. Because it is so, we shall postpone any mention of the latter, other than to say that we shall examine this point sufficiently for our purposes with respect to discrete probability theory in chapter 4. On the other hand, the analogue of arithmetic differences is neatly expressed in set theory as the incarnation of logical negation.

Definition 4. *Let S and T be sets such that $T \subset S$, then we say the complement of T in S is the set consisting of elements in S that are not in T, denoted $S \setminus T$.*

Again, the so-called logical difference defined above can also be described in the ubiquitous elemental notation by $S \setminus T = \{x \,|\, x \in S \text{ AND } x \notin T\}$, where we read the symbol $\notin$ as 'not in,' whence the role of logical negation. These operations for forming new sets from given ones satisfy properties recognizable as properties that the addition and multiplication of integers satisfy. in point of fact, consider the following proposition.

Proposition 1. *Let S, T, and U, be arbitrary sets, then the following statements are true.*

1. *Union of sets is commutative, that is $S \cup T = T \cup S$*
2. *Intersection of sets is commutative, that is $S \cap T = T \cap S$*
3. *Union is associative, that is $(S \cup T) \cup U = S \cup (T \cup U)$*
4. *Intersection is associative, that is $(S \cap T) \cap U = S \cap (T \cap U)$*
5. *Union distributes across intersection, that is $S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$*
6. *Intersection distributes across union, that is $S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$*

We shall prove statement 6 of the proposition as an illustration of the definition of set equality. So, let us show first that $S \cap (T \cup U) \subset (S \cap T) \cup (S \cap U)$. In order to do this, we must show every element of the former set is an element of the latter set. As such, we have, for $x \in S \cap (T \cup U)$ that both $x \in S$ and $x \in T \cup U$. Consequently, $x \in T$ or $x \in U$ and so, $x \in S \cup T$ and $x \in S \cup U$. Conversely, suppose $x \in (S \cap T) \cup (S \cap U)$, then either $x \in S \cap T$ or $x \in S \cap U$, so $x \in S$ and either $x \in T$ or $x \in U$, as desired.

One might notice that, under the analogy between sets and integers, statement 5 is rather unusual, as it would correspond to a distribution law for sums across products. This would mean that, for integers s, t, and u, that $s + t \cdot u = (s + t) \cdot (s + u)$ which is of course, in general, utter nonsense; yet, in this new world or category of sets, one can now easily demonstrate statement 5 as the *dual* of statement 6. We shall not digress to explain exactly what we mean by dual statement beyond indicating it means that if a statement is true of a union and intersection of sets, a corresponding statement is true of the intersection and union of their complements. This powerful way of thinking is used throughout this text and emerges from the following famous precept of logic.

Proposition 2. *Let S and T be subsets of a set Ω, and let $S^c = \Omega \setminus S$. A similar notation is used for T^c, $(S \cap T)^c$, etc. Then the following statements are true.*

1. $(S \cap T)^c = S^c \cup T^c$
2. $(S \cup T)^c = S^c \cap T^c$

Statements 1 and 2 of the proposition are known collectively as *DeMorgan's laws.* We shall not prove these, but one would verify they are true in a way similar to our demonstration of statement 6 above. Moreover, applying DeMorgan's laws to statement 6 implies that statement 5, the so-called dual, is also true.

The following is a digression on notation that is important and useful. All of the above operations can be carried out over indexing sets, which are sets whose elements are intended only to list how other sets are to be substituted into some set theoretic operation. Specifically, let us consider some indexing set I, but most often, we may take $I = \mathbb{N}$ or some finite subset thereof. Then a sequence of sets indexed by I is written as $\{S_i\}_{i \in I}$. The manner in which the various set operations extend to indexed sets is hopefully obvious. For example, we shall consider

$$\bigcup_{i \in I} S_i$$

and say it is the I-fold union of the sets S_i. The usefulness of such general notation is obvious, as often one wishes to appeal to concepts more complex than ones that may only depend upon two or three sets.

Before proceeding to the final concept in this section, let us consider what the additive identity for the logical sum would be according to the analogy we have sketched between set theoretic operations and numerical ones. In other words, what set object would correspond to the number zero? It seems plain that the answer would be the complement of an arbitrary set in itself, as zero is defined as the element satisfying the equation $n - n = 0$ for an arbitrary number $n \in \mathbb{N}$. This discussion motivates the following definition.

Definition 5. *Let S be a set. Then we say $S \setminus S$ is the null or empty set, denoted $\emptyset$*

Indeed, one should notice that the relation $S \cup \emptyset = S$ holds for sets as the relation $n + 0 = n$ holds for integers. Similarly, if we regard the intersection of sets as a kind of product, then the relation $S \cap \emptyset = \emptyset$ holds in sets, as the relation $n \cdot 0 = 0$ also holds in integers.

The final set theoretic concept in this section should be equivalent to fractions or rational numbers with respect to the loose analogy we have developed between sets and integers. However, as mentioned above, this point is nettlesome and will be postponed. Nonetheless, the utility of the concept is immediate and independent of its explanation under the correspondence between operations on sets and operations on integers.

The reason for defining the null set before defining the concept of *partition* below is because it depends on the even more basic concept of *disjointedness*. Indeed, for any sets S and T, we say they are disjoint if their intersection is the null set. This will occur when the overlap of two events in an experiment is logically absurd, and in that context, we shall refer to mutually exclusive events. No matter what, disjoint sets are fundamental throughout this course, as many ideas that follow are statements under the hypothesis, either explicitly or implicitly, that we can use disjoint sets to partition a problem into cells that are easier to examine than initial the set we are given on the whole.

Definition 6. *Let $\{A_i\}_{i \in I}$ be a sequence of subsets of some set S indexed by I. Assume that elements of this sequence are pair-wise disjoint, that is,*

$$A_i \cap A_j = \emptyset$$

for all $i, j \in I$ such that $i \neq j$. Then we say the sequence $\{A_i\}_{i\in I}$ is a partition of S and that the A_i are its cells if

$$\bigcup_{i\in I} A_i = S$$

Moreover, we say the partition is of length r if $|I| = r$.

1.1 WORKED EXAMPLES

1. 1.1 Let $\mathscr{R}$ be the set of all rectangles with integer-valued length and width. Consider $\mathscr{R}(l, w) \subset \mathscr{R}$ the subset of rectangles with either even integer-valued length l or even integer-valued width w. Furthermore, consider $\mathscr{R}(A) \subset \mathscr{R}$ the subset of rectangles whose area is an even integer.

   (a) Prove that $\mathscr{R}(l, w) = \mathscr{R}(A)$

   We proceed by the definition of the equality of sets in this section. That is, $S = T$ if and only if both $S \subset T$ and $T \subset S$. Observe,

   Solution (a) Let $R \in \mathscr{R}(l, w)$, then $Area(R) = l \cdot w$. Since either l or w is an even integer by hypothesis, so is $l \cdot w$. Therefore, $R \in \mathscr{R}(A)$. As R was arbitrary, we have $\mathscr{R}(l, w) \subset \mathscr{R}(A)$. Conversely, let $R' \in \mathscr{R}(A)$. Write $Area(R') = 2m$, then $2m = l \cdot w$. Since 2 is prime, 2 divides either l or w. Therefore, either l or w is even. Thus, as R' was arbitrary, we have $\mathscr{R}(A) \subset \mathscr{R}(l, w)$. Therefore, by the definition of set equality, $\mathscr{R}(l, w) = \mathscr{R}(A)$.

2. 1.1 Simplify the complements of the following sets by DeMorgan's laws.

   (a) $(\{x \geq 5\} \cup \{x \leq 4\})^c$

   (b) Let S, T, U be sets, then simplify $(S \cap T \cup U)^c$

   We proceed by DeMorgan's laws, which indicate how to compose the complement of sets with their unions and intersections.

   Solution (a) We have $(\{x \geq 5\} \cup \{x \leq 4\})^c = \{x \geq 5\}^c \cap \{x \leq 4\}^c$ by DeMorgan's laws. We compute the complement of these intervals by definition, so that
   $\{x \geq 5\}^c \cap \{x \leq 4\}^c = \{x < 5\} \cap \{4 > x\} = (4, 5)$.

   Solution (b) Proceed by both the distribution law and DeMorgans. Observe,
   $S \cap T \cup U = S \cap T \cup S \cap U$ so that
   $(S \cap T \cup U)^c = (S \cap T)^c \cap (S \cap U)^c = S^c \cup T^c \cap S^c \cup U^c = S^c \cup T^c \cap U^c$

3. 1.1 Consider a group of four individuals named *Alan*, *Baker*, *Carmen* and *Duran*. Let S be the set of arrangements of these individuals lined-up in order from first to last. For example, $ABCD$ would represent the order *Alan* is first in line, *Baker* is second in line, etc. Partition S in a manner that depends upon the position of the individuals.

   (a) Partition S in a manner that depends upon the position of the individuals.

   We proceed by definition of a partition of a set S to define cells that depends upon the position of the individuals.

   Solution (a) There are several options available to us according to which position we choose. We shall choose the first position for ease. Accordingly, we may partition S into cells consisting of which of the four individuals are first in the line-up. So define $A_1 =$ *subset of line-ups with Alan first*, $\ldots, A_4 =$*subset of line-ups with Duran first.* Hence $S = \cup_{i=1}^{4} A_i$ since each line-up must be led by an individual and $A_i \cap A_j = \emptyset$ since different individual in first position determine different line-ups.

1.2 Set Functions

In this section we define functions between sets, which should be familiar to the reader as a generalization of real-valued functions introduced and studied carefully in any standard calculus track. Regarding $\mathbb{R}^n$, for $n \geq 0$, as sets, the functions one is familiar with in calculus

$$\begin{aligned} f :&\mathbb{R}^n \to \mathbb{R} \\ &(x_1, x_2, \ldots, x_n) \mapsto y = f(x_1, x_2, \ldots, x_n) \end{aligned}$$

are examples of the following definition.

Definition 7. *Let S and T be sets. The we say a well-defined assignment of elements in S to elements in T is a function, denoted*

$$\begin{aligned} f :&S \to T \\ &s \mapsto t = f(s) \end{aligned}$$

Well-defined in the above definition is intended in the same sense as one intends its use in the aforementioned calculus track. That is, well-defined means that the function f assigns elements in S to unique elements in T. One must be careful as nowadays the word "unique" is abusively used to mean "rare," whereas the mathematician still intends its lexical definition. Moreover, well-defined does not preclude that f may assign several distinct elements of S to the same element of T, for such an assignment is still a unique one with respect to elements of S, which is all our definition requires.

As usual, there is a notion of graph attached to a set function that we shall use to relate experiments to planar regions later on. To describe this, we need another concept from set theory, that of the *Cartesian product of sets.* Given sets S and T, one writes

$S \times T = \{(s,t) | s \in S, t \in T\}$ for the set of ordered pairs of elements in S and T. In particular, the *graph of f* is then the usual subset $\Gamma_f \subset S \times T$ where $\Gamma_f = \{(s,t) | t = f(s)\}$.

The reason in the author's view for all of this formalism is to define several key concepts in this book, least among them being its very name, or at least the word "discrete" in its title. In order to mathematically define this adjective in the present context of set functions, we will explain how the cardinality of various sets are related to each other, for the adjective is one that indicates a particular range of cardinalities for a set.

Of course, in the most elementary approach, one could state a relation between $|S|$ to $|T|$ by simply counting their elements and then comparing the numbers. However, this assumes there are finitely many elements in both sets, and furthermore, even finite, not too many to practically count. Accordingly, this naive approach is insufficient for our purposes even in this modest regard, doubly insofar as we could use such an approach for defining ”discrete.” The following more sophisticated approach will explain what we mean by the word "discrete" and in general clarify what we mean when we attempt to relate the cardinalities of various sets.

Definition 8. *Let $f : S \to T$ be a function between sets S and T. Then we say*

1. *S is the domain of f and T is the codomain or range of f*

2. *For $A \subset S$, one says that $f(A) = \{\, t \,|\, t = f(s) \ \text{for some}\, s \in A\} \subset T$ is the image of A.*

3. *For $B \subset T$, one says that $f^{-1}(B) = \{\, s \,|\, f(s) \in B\,\} \subset S$ is the pre-image of B under f.*

4. *f is injective or 1-1 if $s_1 \neq s_2$ implies $f(s_1) \neq f(s_2)$*

5. *f is surjective or onto if $T = f(S)$*

6. *f is a bijection if it is both injective and surjective, denoted $S \cong T$*

The items in this definition are standard throughout mathematics and are foundational in our subject. For example, later we shall define a partition of a σ-algebra with respect to the equivalence classes induced by the pre-images of a random variable defined on a discrete probability space in order to define the probability of the event "a function equals a number." Such exotic notions require this degree of abstraction to manage. In the present section, let us introduce an important structural feature of sets.

Definition 9. *Let $f : S \to T$ be a function between sets S and T. Then we say the pre-image of $t \in T$ under f is the fibre of f over t.*

The reason we isolate this special case of the definition of pre-image as an important structural feature of sets is because it emphasizes an important conceptual perspective on what it is functions are. Under the auspices of fibres, one may regard the data of a function of sets f as a tool to organize the domain into fibres over single elements in its image. Indeed, the very definition of a function as that of a well-dfined rule mandates this concept, for the fibres over points $t \in f(S)$ partition S. That is,

$$S \cong \bigcup_{t \in f(S)} f^{-1}(t)$$

Different functions with the same domain induce different partitions according to their fibres. As stated already, we introduce this perspective at this earlier stage in the text to anticipate its usage in chapter 4 in the presence of random variables.

In the meantime, we can use the concepts we have defined to answer the question, in general, when $|S| = |T|$. To this end, let us learn of the composition of functions in the generality of sets. Our exposition upon the composition of set functions is the obvious generalization of the composition of real-valued functions. Indeed, let $f : S \to T$ and $g : T \to U$ be a pair of functions, then the function

$$\begin{aligned} g \circ f : & S \to U \\ & s \mapsto u = g \circ f(s) = g(f(s)) \end{aligned}$$

is their composition. This notion allows us to describe both injectivity and surjectivity in terms of functions. To do this, we require the concept of a function's inverse, which, again, is the straightforward generalization of that concept in the context of real-valued functions.

Once more, consider $f : S \to T$, then we say the *left inverse* of f is the function $f_L^{-1} : T \to S$ such that $f_L^{-1} \circ f = Id_S$, where $Id_S : S \to S$ is the identity function of S. One defines a *right inverse* of f *mutatis mutandi.* These concepts motivate the following well-known definition.

Definition 10. *Let $f : S \to T$ be a function. Then we say f^{-1} is the inverse of f if f^{-1} is both the left and right inverse of f.*

The following proposition then expresses the fundamental properties of functions in terms of the existence of inverses.

Proposition 3. *Let $f : S \to T$ be a function. Then the following statements are true.*

1. *f is injective if there exists a left inverse f_L^{-1}*
2. *f is surjective if there exists a right inverse f_R^{-1}*
3. *f is bijective if there exists an inverse f^{-1}*
4. *$|S| = |T|$ if and only if f is a bijection.*

Notice that item 4 of the proposition makes no assumption about the size of either S or T, specifically, whether these sets are infinite or finite. We have been so far vague on this point as technicalities are required to adequately explain, technicalities which are mostly irrelevant in this text beside their relevance to the adjective "discrete," which we can now define.

Definition 11. *Let S be a set. We say S is countable if S there exists a bijection $c : S \to T \subset \mathbb{Z}$ onto a not-necessarily proper subset of the set of integers. Accordingly, there is a distinction between sets in bijective correspondence between proper and improper subsets of $\mathbb{Z}$. In the former case, we say S is finite. In the latter case, we say S is countably infinite.*

Below we shall define discrete probability spaces in terms of this definition. Specifically, they are probability spaces whose sample spaces are countable. In the present, however, this definition may require some context as it is itself rather unusual, for it suggests-famously-the existence of orders of infinity. That is to say, implicitly, it is conceivable there are so-called *uncountable* infinities, and indeed there are. The following theorem is due to Cantor and emphasizes this point.

Theorem 1. *There is no bijection between $\mathbb{Z}$ and $\mathbb{R}$.*

In terms of our definition, one could say, following our italics, that the set of real numbers $\mathbb{R}$ is uncountable, and indeed, this would be correct. The proof of the above theorem is usually proven by Cantor's famous diagonal argument which illustrates there is no bijection between $\mathbb{Q}$ and $\mathbb{R}$. In terms of the theorem, for Cantor's argument to be equivalent, it would follow that $|\mathbb{Z}| = |\mathbb{Q}|$, which is surprisingly true!

Proposition 4. *Let $\mathbb{Q}$ be the set of rational numbers. Then $|\mathbb{Z}| = |\mathbb{Q}|$, that is, $\mathbb{Q}$ is a countable set.*

The proof of this proposition amounts to demonstrating the existence of a bijection between the set of integers and the set of rational numbers. Ideally proving such a seemingly counter-intuitive fact inspires the reader's belief in the necessity of our sophisticated approach to reckoning the cardinalities of sets. So, let us construct such a bijection. First, define $\mathbb{Q}^{\pm}$ to be the sets of positive, respectively negative, rational numbers. Define $c^+ : \mathbb{Q}^+ \to \mathbb{Z}$ by $c^+(\dfrac{m}{n}) = 2^m 3^n$. The fundamental theorem of arithmetic implies that c is injective, so that the image $c^+(\mathbb{Q}^+) \subset \mathbb{Z}$ is a well-ordered subset. Appealing to the well-ordering on the image of c^+, one constructs a second bijection onto $\mathbb{N}$. One recalls that the set of natural numbers is itself countable, thus demonstrating that $\mathbb{Q}^+$ is countable via composition. A similar argument shows that $\mathbb{Q}^-$ is countable. As the union of countable sets is countable and $\mathbb{Q} = \mathbb{Q}^+ \cup \mathbb{Q}^- \cup \{0\}$, we have that the set of rational numbers $\mathbb{Q}$ is countable, as desired.

The proof of this proposition is not entirely self-contained in this text, as it is only intended to illustrate the initial breadth of our new notion of size, that is, reckoning size according to the existence of a bijection. The interested reader is referred to standard introductory texts on Real Analysis. Let us now consider a few examples of the material in this section.

1.2 WORKED EXAMPLES

1. 1.2 Let us prove the following statement from this section. Let S and T be sets, and $f : S \to T$ a set function. Show that $\{f^{-1}(t)\}_{t \in T}$ is a partition of S.

   Solution We must show that the fibres of f are the cells of a partition of S. To this end, we first establish that, for $t \neq t' \in T$, $f^{-1}(t) \cap f^{-1}(t') = \emptyset$. Suppose $s \in f^{-1}(t) \cap f^{-1}(t')$, then by definition of fibre, both $f(s) = t$ and $f(s) = t'$-absurd, as f is well-defined, by hypothesis. Thus, $f^{-1}(t) \cap f^{-1}(t') = \emptyset$. Secondly, $S = \cup_{t \in T} f^{-1}(t)$ since S is the domain of f. Therefore, the fibres of f partition S.

One may interpret this result by thinking of f as dividing S into the, in general unequal portions, of the fibres of f. Later on, viewing set functions in this manner shall play an important role in conceptualizing the role of *random variables* in discrete probability theory.

2. 1.2 Let us now provide an example of the statement in the above section of how different functions between the same sets impart distinct partitions of their shared domain. Consider $S = \mathbb{Z}$ the set of integers and $T = [0, 1, 2, \ldots, 10]$ the set of integers from 0 through 10. Consider the following functions.

   (a) $f : \mathbb{Z} \to [0, 1, 2, \ldots, 10]$ where $m \mapsto r_2(m)$, that is, f maps the integer m to its remainder upon division by 2.

   (b) $g : \mathbb{Z} \to [0, 1, 2, \ldots, 10]$ where $m \mapsto r_3(m)$, that is, g maps the integer m to its remainder upon division by 3.

   Demonstrate that f and g induce distinct partitions of $\mathbb{Z}$.

   Solution (a) To determine the partition $\mathbb{Z}$ induced by f, we merely must consider the image of f, for the partition corresponding to f by the first worked example is indexed by the fibres of f. Therefore, the image of f is self evidently $\{0, 1\} \subset T$, for every integer has either remainder 0 or 1 when divided by 2. Therefore, $\mathbb{Z} = \mathbf{0} \cup \mathbf{1}$, where $\mathbf{0} = f^{-1}(0), \mathbf{1} = f^{-1}(1)$. Notice that the partition induced by f is in bijective correspondence with the set of integers $\{0, 1\}$. Readers familiar with binary notation should recognize this construction.

   Solution (b) Similarly, to determine the partition $\mathbb{Z}$ induced by g, we merely must consider the image of g, for the partition corresponding to g by the first worked example is indexed by the fibres of g. Therefore, the image of g is self evidently $\{0, 1, 2\} \subset T$, for every integer has either remainder 0, 1, or 2 when divided by 3. Therefore, $\mathbb{Z} = \mathbf{0} \cup \mathbf{1} \cup \mathbf{2}$, where $\mathbf{0} = f^{-1}(0), \mathbf{1} = f^{-1}(1)$ and $\mathbf{2} = f^{-1}(2)$. Notice that the partition induced by g is in bijective correspondence with the set of integers $\{0, 1, 2\}$.

   Solution To conclude the example, we note that there is no bijection between the sets $\{0, 1\}$ and $\{0, 1, 2\}$ for their cardinalities are different. As such, f and g impart different partitions to their domain.

3. 1.2 In this example, let us work out the composition of functions together with the partition of its domain.

   (a) Let $S = \{(i, j)\}_{1 \leq i,j \leq 6}$ be the set of pairs of integers from 1 through 6 and $T = \{i + j\}_{1 \leq i,j \leq 6}$ be the set of sums of integers from 1 through 6. Furthermore, let $f : S \to T$ be the obvious addition function.

(b) Let T be as above and $U = [0, 1, \ldots, 5]$ be the integers from 0 through 4. Let $g : T \to U$ be the function that maps $i + j$ to it remainder upon division by 5.

Below, let us first compute the composition of functions and second the partition of the domain of the composition.

Solution The composition of f and g can be obtained by describing its value in terms of an element of its domain. Specifically, $g \circ f((i, j)) = r_5(i + j)$, that is, the remainder after division of 5 into $i + j$ for $1 \leq i, j \leq 6$. Accordingly, $g \circ f : S \to U$.

Solution To compute the fibres of $g \circ f$, we proceed in two steps. First, we compute the fibres of g and and f. Second, we compute the fibre of $g \circ f$ by recognizing it as a union of the fibres f over the fibres of g. Observe, $g^{-1}(0) = \{5, 10\}$ and $f^{-1}(5) = \{(1, 4), (2, 3), (3, 2), (4, 1)\}, f^{-1}(10) = \{(5, 5)\}$. Therefore, $g \circ f^{-1}(0) = \{(1, 4), (2, 3), (3, 2), (4, 1), (5, 5)\}$ or $f^{-1}(5) \cup f^{-1}(10)$. The fibres over other points of the image are computed in a similarly straightforward manner.

1.3 Discrete Probability Theory

In this section we define the first of the three fundamental objects in this text-the other two being that of a discrete random variable and that of a stochastic process. The fundamental object that we define here is the *discrete probability space*. It is a triple consisting of the data (Ω, β, P) described below and subject to certain axioms.

In order to obtain this object from real-world experiments and use it as a model of the likelihood of events that occur in the same, we must first review some basic formal logic. We shall encode logical statements about a real-world experiment in set theory. It is specifically in this manner that one models the likelihood of real-world events in terms of the set theoretic formalism introduced in sections 1 and 2 of this chapter. Specifically, we shall compute the likelihood of the encoded statement.

In this text we shall interpret a real world experiment and events thereof, defined in the informal manner, as equivalent to the set of all logically consistent statements O predicated upon the same. Logical consistency of course means no contradictions are allowed. Accordingly, we are restricted to considering logical statements a and b of the forms $a \wedge b, a \vee b$, and $\overline{a}$, read as a AND b, a OR b, and NOT a, respectively. The conjunction of statements and the negative or "not" statement are true according to the truth values of the statements in conjunction. We summarize the truth values of these conjunctions in the following tables.

$$\begin{bmatrix} \text{AND} & a & \wedge & b \\ & \text{T} & \text{T} & \text{T} \\ & \text{F} & \text{F} & \text{T} \\ & \text{T} & \text{F} & \text{F} \\ & \text{F} & \text{F} & \text{F} \end{bmatrix}$$

$$\begin{bmatrix} \text{OR} & a & \vee & b \\ & \text{T} & \text{T} & \text{T} \\ & \text{F} & \text{T} & \text{T} \\ & \text{T} & \text{T} & \text{F} \\ & \text{F} & \text{F} & \text{F} \end{bmatrix}$$

and $\overline{a}$ is F if and only if $a = \text{T}$. DeMorgan's laws can be expressed in this notation by $\overline{a \wedge b} = \overline{a} \vee \overline{b}$ and $\overline{a \vee b} = \overline{a} \wedge \overline{b}$. Notice that negation is idempotent, or $\overline{\overline{a}} = a$. We say a logical statement is *simple* if it cannot be expressed as the either $\wedge$ or $\vee$ of distinct, non-trivial statements.

We want to relate the set O of such logical statements about an experiment to a set, say Ω, to translate actual experiments into mathematics. In general, we say a set A *verifies* a logical statement α if there exists $a \in A$ such that $\alpha(a) = T$ or α predicated upon a is a true statement.

Definition 12. *Let O be all the simple logical statements about an aspect or aspects of a real-world experiment germane to an experimenter. Then we say the set that verify the logical statements in O is the sample space of the experiment, denoted Ω. Moreover, we say any subset of Ω is an event.*

Under the mapping from logical statements about a real-world statement to the sample space Ω induced by this definition, we obtain another map from the set of all conjunctions, disjunctions, and negations of statements in O and the set of all subsets of Ω, say β, given by

$$\begin{aligned} a \wedge b &\mapsto A \cap B \\ a \vee b &\mapsto A \cup B \\ \overline{a} &\mapsto A^c \end{aligned}$$

where $A, B \subset \Omega$ are the events that verify the statements a, b, respectively. The crucial insight is that we can now assess the likelihood of some logical statement a about some real-world experiment in terms of the event A that verifies it. As the latter is an object of mathematics, one could then say the above correspondence is tantamount to a mathematical model of the real-world experiment. However, at this stage, this model would be incomplete. The incompleteness of this putative model is the motivation to include the minimum of data sufficient to rigorously define the likelihood of a statement or the *probability of the corresponding event.*

Let us continue now by introducing an intuitive, although insufficient, definition of the probability of an event. We intuitively define the probability of a subset of the sample space to be its value under a set function $P : \Omega \to [0, 1]$, where $[0, 1]$ is the unit interval in the set of real numbers. Namely, for $A \subset \Omega$, one has $0 \leq P(A) \leq 1$ or $P(A) \in [0, 1]$, imparting rigor to our intuition that the likelihood of a statement about an experiment is some value between 0 %-likely and 100 %-likely. Although creative, this definition requires refinement as the set function P does not, in general, preserve set operations, so that, for example, $P(A \cup B) \neq P(A) \cup P(B)$. As such, one would be unable to compute the probability of set theoretic operations applied to events, which would necessarily be a requirement in general if we are to make sense of the likelihood of an arbitrary statement in our model. The problem with this intuition is the domain of P, namely,the sample space Ω. We rectify this insufficiency in our intuition in the following way.

Definition 13. *Let S be a set. Then we define the power set of S to be the set of all subsets of S, denoted 2^S.*

Observe that replacing the sample space Ω by its power set 2^Ω, denoted hereafter by β, in our intuitive approach above, obviates the problem it had encountered, for one does not need to preserve set operations in order to evaluate the probability of set theoretic combinations of events. For example, if

$$P : \beta \to [0, 1]$$

then the probability of the event $A \cup B \in \beta$ simply *is* $P(A \cup B)$, as no appeal to the values under P of the individual sets in this union is required. So, refining the data of the correspondence between statements a about real-world experiments and events $A \in \beta$ that verify them given above entails the following definition.

Definition 14. *Let Ω be the sample space of a real-world experiment and β its power set, then we say any set function $P : \beta \to [0, 1]$ such that following axioms are satisfied is a probability measure or probability function on Ω.*

1. $P(\Omega) = 1$

2. $P(A) \geq 0$ *for any* $A \in \beta$

3. *Let $\{A_i\}_{i \in I}$ be a sequence of events in Ω that are pair-wise disjoint, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$P(\bigcup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$$

The three axioms in this definition are quite famous and should be committed to memory. They are known as *Kolmogorov's axioms* and we shall have many occassions throughout this text to appeal to the structure they impart to problems and concepts.

So far in our discussion we have involved, in no substantial way, any hypothesis that distinguishes our subject-that of *discrete* probability theory. In turns out that, in the generality we have worked with so far, assuming $\beta = 2^\Omega$ would be a mistake. Rather, one must assume that β has the structure of a so-called σ-algebra, defined as follows.

Definition 15. *Let S be a set. Then we say a set of subsets β of S such that the following closure axioms are satisfied is a σ-algebra.*

1. *β is closed under complements, that is, if $A \in \beta$, then $A^c \in \beta$.*

2. *β is closed under the empty set, that is $\emptyset \in \beta$*

3. *β is closed under arbitrary unions, that is, if $\{A_i\}_{i \in I} \in \beta$, then $\cup_{i \in I} A_i \in \beta$*

DeMorgan's laws imply that, for any σ-algebra β, we also have $S \in \beta$ for $S^c = \emptyset$ and β is closed under complements. Similarly, β is closed under arbitrary intersections, for if $\cup_{i \in I} A_i \in \beta$ then so is $(\cup_{i \in I} A_i^c)^c = \cap_{i \in I} A_i$, by DeMorgan's laws. Taken together, for reasons beyond the scope of this book, β has the structure of a mathematical object called an algebra, whence its name. It is fascinating to reflect on the fact our logical impressions of a real-world experiment have an algebraic structure. Putting that matter to the side, we have the following proposition.

Proposition 5. *Let S be a countable set, then its power set 2^S is a σ-algebra*

We shall not prove this proposition, but at least we know for sure how to define the probability of an event when a sample space is countable. Equivalently, that there is a canonical definition for β whenever Ω is countable. This motivates the first definition of the three fundamental objects in this text and summarizes the above discussion for the remainder of this text.

Definition 16. *Let Ω be a countable sample space of a real-world experiment, then we say any triple (Ω, β, P) where $\beta = 2^\Omega$ is the canonical σ-algebra associated to Ω, and P is a probability measure satisfying Kolomogorov's axioms, is a discrete probability space. Moreover, we say the probability of an event $A \in \beta$ is its value under P.*

So, with this definition at hand, one defines discrete probability theory as the aspect of probability theory germane to discrete probability spaces. Specifically, to studying the likelihood of statements about experiments with but countably many logically simple statements.

The canonical nature of a discrete probability space extends even unto P itself, our axiomatization of the likelihood of a statement. The canonical definition of P is given in terms of Ω by the following theorem.

Theorem 2. *Let (Ω, β, P) be a discrete probability sample space such that $|\Omega| = n$ is finite. In terms of its simple events, write $\Omega = \{u_1, u_2, \dots, u_n\}$. Then for any set of real numbers $p_i = P(u_i) \in [0, 1]$ for $1 \leq i \leq n$ such that $\sum_{i=1}^n p_i = 1$, and arbitrary event $A \in \beta$, the probability function*

$$P(A) = \sum_{i \,|\, u_i \in A} p_i$$

satisfies Kolmogorov's axioms.

The proof of this theorem is simply a matter of verifying that the definition of P satifies Kolomogorov's axioms given the hypotheses. So, for $A \in \beta$, it is obvious that $P(A) \geq 0$ since $p_i \geq$ for each $u_i \in A$. As $\sum_{i=1}^n p_i = 1$ and $\Omega = \cup_{i=1}^n u_i$ it follows that $P(\Omega) = 1$. Lastly, for $\{A_1, A_2, \ldots, A_k\}$ such that the A_i are pair-wise disjoint, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^{k} A_i\right) &= P\left(\sum_{j|u_j \in \cup_{i=1}^{k} A_i} p_j\right) \\ &= \sum_{i=1}^{k}\left(\sum_{j|u_j \in A_i} p_i\right) \\ &= \sum_{i=1}^{k} P(A_i) \end{aligned}$$

where the middle equality is obtained by the hypothesis that the A_i are pair-wise disjoint. This remark concludes the proof of the theorem. We conclude with the remark that, more generally, the statement of the theorem is true for countable Ω given a sequence of real numbers $\{p_i\}_{i=1}^{\infty}$ such that corresponding series $\sum_i p_i$ converges to 1.

1.3 WORKED EXAMPLES

1. 1.3 Let us consider an example of an abstract discrete probability space, which is to say, a discrete probability space for which there is no experimental significance for its events. To wit, let $\Omega = \{A, B, C\}$ be the set of simple events so that $\beta = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, \Omega\}$. The values $p_1 = .25, p_2 = .5$, and $p_3 = .25$ satisfy Kolmogorov's axioms when assigned to A, B and C under P, respectively. In this manner, $P(A \cup B) = P(A) + P(B) = .25 + .5 = .75$. The significance of this example is that the probability of events can be furnished by an assignment consistent with Kolmogorov's axioms-there does not need to be any experimental justification for the assignment, as one might anticipate intuitively.

2. 1.3 Consider the experiment of tossing a fair coin and recording the outcome. The sample space associated with this experiment consists of the simple events $\Omega = \{H, T\}$.

   (a) Assuming $P(H) = P(T)$, use Kolmogorov's axioms to find $P(H)$.

   Solution (a) Observe that, if Kolomogorov's axioms hold, we have both $P(\Omega) = 1$ and $\Omega = T \cup H$, so that $1 = P(H \cup T) = P(H) + P(T)$. The hypothesis assumes $P(H) = P(T)$, so therefore $1 = 2P(H)$ or $P(H) = \frac{1}{2}$.

3. 1.3 Consider the experiment of tossing a fair coin three times and recording the number of heads.

   (a) Describe the sample space associated to this experiment.

   (b) Assign probabilities to the simple events identified in (a) that are consistent with Kolmogorov's axioms.

(c) Compute the probabilities of the events $A = \{1, 2, 3\}$ and $B = \{0, 3\}$.

Solution (a) Proceed by the definition of the sample space by listing the simple events associated with this experiment. Since we are recording the number of heads which might appear on the faces of three tossed coins, the range of such is the simple events. Thus, with obvious notation, we have $\Omega = \{0, 1, 2, 3\}$.

Solution (b) As there are four simple events, we may choose any p_i satisfying the condition that their sum is 1, for $i = 1, 2, 3, 4$. However, let us select values for these that accord with intuition, especially since the experiment is described recording the results of fair coins. As such, for $P(i-1) = p_i$, we have $p_1 = \frac{1}{8}, p_2 = \frac{3}{8}, p_3 = \frac{3}{8}$ and $p_4 = \frac{1}{8}$.

Solution (c) As simple events are mutually disjoint, we have by Kolmogorov's third axiom that $P(A) = P(1) + P(2) + P(3) = p_2 + p_3 + p_4$ and, similarly, $P(B) = P(0) + P(3) = p_1 + p_4$.

4. 1.3 Consider the experiment of a horse race in which three horses, labeled 1,2 and 3, are meant to race, and recording the winner.

   (a) Assume 1 is twice as likely to win as 2 and 2 is twice as likely to win as 3. Compute the probabilities of each horse winning the race.

   (b) Compute the probability that either horse 1 or horse 2 wins the race.

Solution (a) Let us abusively denote the simple events that a horse wins by their label, so that the sample space associated to this experiment can be written as $\Omega = \{1, 2, 3\}$. As such, denote by p the probability $P(3)$ or of the event, "horse 3 wins". By hypothesis, $P(2) = 2p$ and $P(1) = 2(2p) = 4p$. By Kolmogorov's second axiom, $P(\Omega) = 1 = 4p + 2p + p = 7p$ and therefore, $p = \frac{1}{7}$. Substituting this value for p finishes the work.

Solution (b) $P(1 \cup 2) = P(1) + P(2)$ by Kolmogorov's third axiom, which in turn is given by the sum $\frac{4}{7} + \frac{2}{7} = \frac{6}{7}$.

1.4 The Relative Frequency of an Event and Examples of Discrete Probability Spaces

In this section we introduce the first of two fundamental applications of the theorem canonically defining the value of $P(A)$ for some event $A \in \beta$ in a discrete probability space.

The second application is to computing $P(A)$ for some event in an experiment that occurs over time, which speaks to the definition of an event within a stochastic process. However, in the present context, we study the first application which is to compute the probability of an event A that is observed in an experiment that is performed repeatedly. The interpretation of the value $P(A)$ is often referred to as the *relative frequency of the event* A and best reflects one's intuitive understanding of what it means to compute the probability of an event A to occur.

Suppose an experiment may be repeated a finite number of times, say n, in exactly the same manner and that each repetition of the experiment, together with its outcome, is independent of the prior performance. Suppose further a is some logical statement about the n repetitions of the experiment, then the likelihood of the statement a is equivalent to the probability of the corresponding event A that verifies it. Indeed, one refers to $P(A)$ as the *relative frequency of* a and computes its value, intuitively, to be the number of times it had occurred as portion of the n performances of the experiment. One regards each performance of the experiment as a simple event in terms of the corresponding discrete probability space, so that one computes $P(A)$ to be the number of performances or simple events favorable to the outcome of a divided by the number of performances of the experiment. Specifically,

$$P(A) = \frac{|A|}{|\Omega|}$$

This interpretation justifies our introduction to combinatorics in the second chapter of this text, for to compute the probability of the event A according to its relative frequency is tantamount to *counting* the cardinalities of both A and Ω. Furthermore, we must observe that the intuitive argument is subsumed by the axiomatic treatment of probability in section 1.3. Indeed, it suggests an important class of spaces in their own right, that of *equiprobable spaces*, that is, discrete probability spaces such that the probability of each simple event is a constant $0 \leq p \leq 1$. In particular, the relative frequency interpretation of the probability of an event entails a discrete equiprobably space such that the probability of each simple event is $p = \dfrac{1}{n}$, for the portion of each performance of the experiment or simple event is one of the n performances. Consequently, applying the canonical definition of P to an arbitrary event in a discrete equiprobable space, we obtain

$$P(A) = \sum_{u_i \,|\, u_i \in A} \frac{1}{n}$$

which is equivalent to the relative frequency interpretation for $|A| = |\{u_i|\, u_i \in A\}|$, so that $P(A) = \dfrac{|A|}{|\Omega|} = \displaystyle\sum_{u_i \,|\, u_i \in A} \frac{1}{n}$.

The above discussion is nothing more than a special case of the theorem of 1.3. As such, it establishes that the relative frequency interpretation of the probability of an event is a special case of the axiomatic interpretation. We have occasion to adopt this general view of what probability *is* in order to discuss cogently what the probability of an event in an experiment that occurs over time, that is, the second fundamental application of the major application of probability theory in this text. In the meantime, we shall consider some examples, primarily to demonstrate the breadth that the axiomatic interpretation of probability affords us and the scope of the work the relative frequency interpretation motivates.

1.4 WORKED EXAMPLES

1. 1.4 Determine the probability $P(A) = p$ of each of the following events:

   (a) $A =$ *A king appears in the drawing of a single card from an ordinary deck of 52 cards*

   (b) $A =$ *At least one tail appears in the toss of three fair coins*

   (c) $A =$ *A white marble appears in drawing a single marble from an urn containing 4 white marbles, 3 red marbles, and 5 blue marbles*

   We proceed by computing the relative frequency of each of these events to find p

   Solution (a) A king appears in 4 ways when drawing a single card from an ordinary deck of 52 cards. Therefore, $p = \dfrac{4}{52}$.

   Solution (b) A least one tail appears in the toss of three fair coins, represented by triples $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$, in exactly 7 simple events. Therefore $p = \dfrac{7}{8}$.

   Solution (c) A white marble appears in drawing a single marble from an urn containing 4 white marbles, 3 red marbles, and 5 blue marbles in 4 ways. Therefore, $p = \dfrac{4}{12}$.

2. 1.4 Two cards are drawn at random from an ordinary deck of 52 cards. Compute the probabilities of the following events:

   (a) $A =$ *both cards drawn are spades*

   (b) $B =$ *one is a spade and one is a heart*

   We proceed by computing the relative frequency of each of these events.

   Solution (a) To compute $P(A)$, we must compute $\dfrac{|A|}{|\Omega|}$. In particular, $|\Omega|$, which is the number of 2 card draws from a deck of 52 is 1326. We shall prove this fact later in chapter by recognizing such pairs as 2-combinations of the deck. Similarly, there are 78 ways to draw two spades from the subset or suite of spades in a 52 card deck. Therefore, $P(A) = \dfrac{78}{1326}$.

Solution (b) To compute $P(B)$, we must compute $\dfrac{|B|}{|\Omega|}$. We observe that $|\Omega| = 1326$ as in the previous solution for the same reason. As for $|B|$, there are 169 ways to draw a spade and a heart. We shall prove this below in chapter 2 after we learn the multiplication rule. Therefore, $P(B) = \dfrac{169}{1326}$.

3. 1.4 Twelve individuals standing together in pairs are standing together in a room. Compute the probabilities of the following events:

   (a) $A =$ *If two individuals are chosen at random, find the probability that they were originally together in a pair*

   (b) $B =$ *If four individuals are chosen at random, find the probability that both couples were originally together in a pair*

   (c) $C =$ *If four individuals are chosen at random, find the probability that exactly one couple were originally together in a pair*

We proceed by computing the relative frequency of each of these events.

Solution (a) To compute $P(A)$, we must compute $\dfrac{|A|}{|\Omega|}$. First, we observe there are 66 ways to pair 12 individuals. We shall prove this later in chapter 2 by recognizing 66 as the number of 2-combinations of a set of 12 elements. Now there are 6 pairs of twelve people standing together in pairs. Therefore, $P(A) = \dfrac{6}{66}$.

Solution (b) To compute $P(B)$, we must compute $\dfrac{|B|}{|\Omega|}$. First, we observe there are 495 ways to choose 4 individuals from 12. We shall prove this later in chapter 2 by recognizing 495 as the number of 4-combinations of a set of 12 elements. Now there are 15 ways to choose 2 pairs from 6 pairs. Again, we shall show why this is true below, however, for the time being, we take these results as motivation to learn combinatorics to compute the probability of events in a discrete probability space. Therefore, $P(B) = \dfrac{15}{495}$.

Solution (c) To compute $P(C)$, we must compute $\dfrac{|C|}{|\Omega|}$. First, $|\Omega| = 495$ for the same reason as in the previous solution. To compute $|C|$, we recognize C can be described as either both couples were originally in pairs or neither were. The first half of this description has probability $\dfrac{15}{495}$ by the previous exercise. The second half of this description has probability $\dfrac{240}{495}$. The reason is because there are 240 ways to choose two couples where no member of the couple was originally standing together in a pair. This follows from the following argument. There are 15 ways to choose 4 people from 6 couples, and since we insist they were not originally standing together in a pair, we must choose one of the 2 from each of the 4 pairs standing together originally. Therefore, $P(C) = \frac{15}{495} + \frac{240}{495}$.

1.5 The Calculus of Discrete Probability Spaces

In this final section of chapter 1 we prove some basic and widely known results establishing the calculus involved in computing the probabilities of arbitrary events. We conclude with the statement and proof of the law of total probability, a result that looms large over the remainder of the results in this text. In particular, we require it to prove the famous Baye's theorem in chapter 3 and it plays a fundamental role in proving that the transition probabilities between states in a Markov chain satisfy the fundamental axiom distinguishes Markov processes from stochastic ones.

Proposition 6. *Let (Ω, β, P) be a discrete probability space and $A \in \beta$ an arbitrary event. The following statements are true.*

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$

The proofs of these statements are elementary. Notice that $\Omega = A \cup A^c$ for any event and that since P satisfies Kolmogorov's axioms, we have

$$\begin{aligned} P(\Omega) =& P(A \cup A^c) \\ 1 =& P(A) + P(A^c) \end{aligned}$$

which gives both statements 3 and 1 immediately-the latter by taking the special case when $A = \Omega$. As for statement 2, we know that $1 = P(A) + P(A^c)$ because 1 is true, so that $P(A) \leq 1$. Next, we consider the probability of the union of arbitrary events in the following proposition. It should be obvious why such a computation would be of interest, for events whose probabilities are not a consequence of Kolmogorov's axioms will in general be those that will occur in applications.

Proposition 7. *Let (Ω, β, P) and $A, B \in \beta$ be arbitrary events. The following statements are true.*

1. $P(A^c \cap B) = P(B) - P(A \cap B)$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. *Assume $A \subset B$, then $P(A) \leq P(B)$, that is, P is an inclusion preserving function.*

The trick to this proof is to recognize that can one decompose an arbitrary event B relative to A as

$$B = (B \cap A^c) \cup (B \cap A)$$

After applying P, this decomposition generalizes the computation of the axioms, for there one takess $A = \Omega$ to compute what one might say is the *absolute* probability of B in this context. Specifically, under that substitution, one has $B = (B \cap \emptyset) \cup (B \cap \Omega)$ which tautologically is B. Hence $P(B) = P(B \cap \emptyset) \cup (B \cap \Omega)$. Thus, in this generality, one might say we are considering the *relative* probability of B with respect to A.

Given the relative set theoretic description of an event B with respect to A, the proofs of the three statements become more or less obvious. Statement 1 is true because $B \cap A$ and $B \cap A^c$ are mutually exclusive events, so 1 follows from Kolmogorov's third axiom. Statement 2 is true because $A \cup B = A \cup (B \cap A^c)$, so that, as the events on the right hand side are mutally exclusive, we have $P(A \cup B) = P(A) + P(B \cap A^c)$, but then

$$\begin{aligned} P(A \cup B) =& P(A) + P(B \cap A^c) \\ =& P(A) + P(B) - P(A \cap B) \end{aligned}$$

by substituting statement 1.

Last, to show statement 3 is not difficult, either, given the hypothesis that $A \subset B$. Indeed, it then follows from statement 1, since $P(B \cap A^c) = P(B) - P(B \cap A) \geq 0$ by Kolmogorov's axiom 1. Statement 1 then simplifies to $P(B) - P(A) \geq 0$ by the hypothesis, for $A \cap B = A$ when $A \subset B$. The rest of the argument is obvious.

A well-known corollary of the proposition is Bonferroni's Inequality. Indeed, by statement 2, we have $P(A \cup B) \leq 1$ which gives that $P(A) + P(B) - P(A \cap B) \leq 1$ or

$$P(A \cap B) \geq P(A) + P(B) - 1$$

which is the inequality.

Although interesting, statement 2 is not the most general, as it leaves open the question, "what is the probability of a finite union of arbitrary events, say $A_1, A_2, \ldots A_k$?" We shall not pursue the proof of the answer to this question, but merely state that it is beyond the scope of this text. Nonetheless, the combinatorial nature of this proof lends itself to naming this theorem, which usually is known as the probabilistic principle of inclusion-exclusion.

Theorem 3. *Let (Ω, β, P) be a discrete probability space and $A_1, A_2, \ldots, A_k \in \beta$ be a sequence of arbitrary events, then*

$$P(\cup_{i=1}^k A_i) = \\ \textstyle\sum_{i=1}^k P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<l} P(A_i \cup A_j \cup A_l) + \cdots + (-1)^{k-1} \sum_{i<\ldots<l} P(\cap_{i=1}^k A_i)$$

One has as a direct application of the inclusion-exclusion principle Boole's Inequality, which simply states that, with hypotheses and notation as above, that

$$P(\cup_{i=1}^k A_i) \leq \textstyle\sum_{i=1}^k P(A_i)$$

One has an equality whenever $A_1, A_2, \ldots, A_k$ is a sequence of pairwise disjoint events, according to Kolmogorov's axiom 3. Thus, Boole's inequality addresses the general case.

Finally, we conclude both this section and chapter with the law of total probability, a result whose implications we shall explore both explicitly and implicitly for the remainder of this text. Again, we hope to indicate now its relevance to Markov processes and Markov chains below in order to justify its importance to the reader.

Definition 17. *Let (Ω, β, P) be a discrete probability space, then we say a sequence of events $\{A_i\}_{i \in I} \in \beta$ that are pairwise-disjoint and that cover Ω in the sense that*

$$\Omega = \cup_{i \in I} A_i$$

is a partition of Ω. We refer to the elements of the sequence $\{A_i\}_{i \in I}$ as the cells of the partition. Furthermore, when $|I| = r$ for some finite natural number r, we say $A_1, A_2, \ldots, A_r$ is a partition of length r.

A partition of a discrete probability space helps us to compute the probability of an arbitrary event in terms of its restriction to the cells of the partitions. As this result applies to an arbitrary event B, one refers to the following theorem as *the law of total probability.*

Theorem 4. *Let (Ω, β, P) be a discrete probability space together with a partition of length r, say $A_1, A_2, \ldots, A_r$. Then for an arbitrary event $B \in \beta$, we have*

$$P(B) = \sum_{i=1}^{r} P(B \cap A_i)$$

The proof relies upon the relative view of the probability of an event B with respect to the cells of the partition. Indeed, by hypothesis, the partition of length r covers Ω, so that we can write

$$\begin{aligned} B =& B \cap \Omega \\ =& B \cap (\cup_{i=1}^{r} A_i) \\ =& \cup_{i=1}^{r} (B \cap A_i) \end{aligned}$$

where the events $A_i \cap B$ are pairwise-disjoint by the fact that the A_i are the cells of a partition. Therefore, by Kolmogorov's third axiom, we have

$$\begin{aligned} P(B) =& P(\cup_{i=1}^{r} (B \cap A_i)) \\ =& \sum_{i=1}^{r} P(B \cap A_i) \end{aligned}$$

as desired.

1.5 WORKED EXAMPLES

1. 1.5 Consider the experiment that consists of predicting the order made in a restaurant. Suppose there are three meals that ca be ordered: A for a meal of apples, B for a meal of beans, and C for a meal of carrots. Suppose further that, with obvious notation, $P(A) = .7, P(B) = .8, P(C) = .75, P(A \cup B) = .85, P(A \cup C) = .9, P(B \cup C) = .95$, and $P(A \cup B \cup C) = .98$.

   (a) What is the probability that order is for at least one of apples, beans, or carrots?

   (b) What is the probability the diner will order none of apples, beans, or carrots?

(c) What is the probability the diner will only order apples and neither beans nor carrots?

(d) What is the probability the diner will only order one of the three meals, apples, beans, or carrots?

Solution (a) We must compute $P(A \cup B \cup C)$ but this is given to us in the question, so $P(A \cup B \cup C) = .98$

Solution (b) We must compute $P((A \cup B \cup C)^c)$. Therefore, $P((A \cup B \cup C)^c) = 1 - .98 = .02$

Solution (c) There are several steps in this computation, as we must excise the subset of events consisting of apples and beans, apples and carrots, and apples, beans, and carrots. First, we compute $P(A \cap B) = P(A) + P(B) - P(A \cup B) = .65$ after substituting the values in our hypothesis. Next, we compute both $P(A \cap C) = .55$ and $P(B \cap C) = .6$ in a similar manner. Last, we compute $P(A \cap B \cap C) = P(A \cup B \cup C) - P(A) - P(B) - P(C) + P(A \cap B) + P(A \cap C) + P(B \cap C)$ in the obvious manner extending the first two computations. Substituting the values from the hypothesis and our work so far, we have $P(A \cap B \cap C) = .53$ and therefore, $P(\text{onlyA}) = .7 - (.55 - .53) - (.65 - .53) - .53 = .03$. We remark that a Venn diagram may help one to keep track of the computation.

Solution (d) Last, we must compute $P(\text{onlyA} \cup \text{onlyB} \cup \text{onlyC})$. The probabilities of the events onlyB and onlyC are computed in a manner similar to the computation of $P(\text{onlyA})$. Therefore, $P(\text{onlyA} \cup \text{onlyB} \cup \text{onlyC}) = P(\text{onlyA}) + P(\text{onlyB}) + P(\text{onlyC})$ since these events are mutually disjoint, and so, $.03 + .08 + .13 = .24$.

2. 1.5 Consider the experiment of recording whether a passerby trips, falls, or rolls. Suppose in observing a passerby, the probabilities assigned to events in this experiment are, with obvious notation, are $P(T) = .12, P(F) = .07, P(R) = .05, P(T \cup F) = .13, P(F \cup R) = .14, P(F \cup R) = .1$, and $P(T \cup F \cup R) = .01$.

   (a) What is the probability that a passerby does not trip?

   (b) What is the probability a passerby both trips and falls?

   (c) What is the probability a passerby both trips and falls but does not roll?

   (d) What is the probability a passerby experiences at most two of the events, e.g. trip and fall?

Solution (a) Compute $P(T^c) = 1 - P(T) = 1 - .12 = .88$ by the complement rule.

Solution (b) We must compute
$P(T \cap F) = P(T) + P(F) - P(T \cup F) = .12 + .07 - .13 = .06$

Solution (c) We must compute $P(T \cap F \cap R^c)$ to solve the problem, for we want to know the probability of the event that the passerby trips AND falls AND does not roll. A Venn diagram should convince the reader that $T \cap F \cap R^c = T \cap F \setminus T \cap F \cap R$ so that $P(T \cap F \cap R^c) = P(T \cap F) - P(T \cap F \cap R) = .06 - .01 = .05$. Notice that we leave the computation of $P(T \cap F \cap R)$ to the reader, for both it and the required probabilities of intersections of events are computed in a manner similar to the computations made in the previous example.

Solution (d) The probability of this last event is obtained by recognizing that probability of at most two of the events is the complement of the probability of all three occurring. Therefore, by the complement rule, we have
$P(\text{atmost2}) = 1 - P(T \cap F \cap R) = 1 - .01$ by the previous solution.

1.6 Chapter 1 Homework Exercises

1. Allen, Baker, Cabot, and Dean are to speak at a dinner. They will draw lots to determine the order in which they will speak. Please answer the following questions:
   a.) List all the elements of a sample space Ω associated to the experiment of recording the order in which these four individuals speak at the dinner.

   b.) Mark with a check the simple events in part a.) contained in the event $A = \{\text{Allen speaks before Cabot}\} \subset \Omega$

   c.) Mark with a cross the elements of the event, $A = \{\text{Cabot's speech is between those of Allen and B}$

   d.) Mark with a star the elements of the event $A = \{\text{The four persons speak in alphabetical order}\}$

2. Consider the experiment "A coin is tossed five times."

   a.) Determine the sample space Ω associated to this experiment. Furthermore, count the number of simple events in the same. In set theoretic notation, this means to compute $|\Omega|$ i.e. the cardinality of the sample space as a set.

   Please use the fact that the discrete probability space (Ω, β, P) determined by the experiment in a.) is an equally likely sample space to compute the following probabilities of the events named in the following questions:

   b.) Heads never occurs twice in a row.

c.) Neither heads nor tails ever occurs twice in a row.

d.) Both heads and tails occur at least twice in a row.

3. Consider the experiment "Two dice are thrown." the sample space Ω in our class notes. Let

$$\begin{aligned} A &= \{\text{The total is two}\} \\ B &= \{\text{The total is seven}\} \\ C &= \{\text{The number shown on the first die is odd}\} \\ D &= \{\text{The number shown on the second die is odd}\} \\ E &= \{\text{The total is odd}\} \end{aligned}$$

be events in this experiment. Given that $(\Omega, \mathscr{B}, P)$ is an equally likely sample space, compute the following probabilities:
a.) $P(A)$

b.) $P(B)$

c.) $P(C)$

d.) $P(D)$

e.) $P(E)$

f.) $P(A \cup B)$

g.) $P(A \cap B)$

h.) $P(A \cup C)$

i.) $P(C \cap D \cap E)$

j.) $P(B \cup D^c)$

4. Prove, by the aid of Venn diagrams, that for any probability space (Ω, β, P) and events $A, B, C \in \beta$ that

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

Recall the following definition in set theory.

Definition 18. *Let A and B be sets, then we say*

$$A \setminus B = \{x \in A | x \in A, x \notin B\}$$

is the difference of B in A.

Remark: "by aid of Venn diagrams" means drawing a set of Venn diagrams that exhibits the underlying set theoretic relation in conjunction with a remark to that effect that, "because P is a probability distribution function it does this by the axioms of a probability space" is sufficient. TLDR: you can just use pictures.

5. Recall the defintion of the **power set** of an arbitrary set, A: We say the set of all subsets of A, denoted 2^A, is the power set of A.
Let $A = \{a, b, c\}$ be a set of three elements. Consider the set function

$$f : 2^A \to \mathbb{Z}$$

defined by $f(B) = |B|$ that is, the image of a subset of A under f is the number of its elements.
Draw a diagram that illustrates this function that includes both its domain and its image. Use this diagram in conjunction with the definition to illustrate why f is not an injection i.e. is not 1-1.

6. Recall, a rule assigning elements of a set A to a set B is said to be **well-defined** if the assignment of elements in A to those of B is unique. Furthermore, such a rule is said to be a function when it is well-defined. Consider the rule

$$f : \mathbb{Q} \to \mathbb{Z}$$

where $f\left(\frac{m}{n}\right) = m - n$ for any $q = \frac{m}{n} \in \mathbb{Q}$. Explain why this rule is not a function.

7. Suppose A, B, C partition the set Ω and that (Ω, β, P) is a probability space. Suppose further $P(A) = 0.3$ and $P(B) = 0.5$. Compute both the probability $P(A \cup B)$ and the probability $P(C)$.

8. Suppose $A, B \in \beta$ are events in the probability space (Ω, β, P) and $P(A) = 0.8, P(B) = 0.7$, and $P(A \cap B) = 0.6$. What is the probability $P(A \cup B)$?

2 Arrangements and Elementary Combinatorics

In this chapter we study the introduction to counting ideas and concepts required to compute the probability of events with respect to the relative frequency interpretation of probability discussed in section 1.4. Indeed, the most well-known interpretation of the probability of an event is to count the number of performances of an experiment favorable to it and to divide this figure by the total number of performances of the experiment-both figures require counting techniques germane to counting the size of various sets, collectively known in mathematics as combinatorics. Moreover, after we have introduced discrete random variables in chapter 4, we shall apply the counting techniques of this chapter to simplifying families of probabilities spaces by obtaining parameters that characterize the probabilities of their events. In particular, such parameterizations worked out in chapter 5 shall depend upon counting what we call arrangements in this chapter. Accordingly, we restrict the scope of our study of combinatorics to studying the number of various such arrangements of the elements of a set.

2.1 The Multiplication Rule

In this section we state and prove the multiplication, which is a counting technique that shall underlie all of the combinatorics in this text. It is for this reason we assert that our treatment of this subject is elementary. We begin with a simple counting lemma we shall use several times below.

Lemma 1. *Let $r, n \in \mathbb{N}$ such that $r \leq n$, then the number of integers from r through n is $n - r + 1$*

The formal proof is by induction and it is not difficult. Informally, one must keep track of the position of the integers and count the total number of positions. Observe that, in the following list

$$\begin{aligned} r &= r + 0 \\ r + 1 &= r + 1 \\ &\dots \\ r + i &= r + i \\ &\dots \\ n &= r + (n - r) \end{aligned}$$

i keeps track of the position of an integer from r through n. Therefore, $r + i$ is in the $1 + i$-th position. In particular, there are $n - r + 1$ integers from r through n.

As an application, consider counting the number of integers from 120 through 999 divisible by 5. One may write every fifth number in this range as as divisible by 5, so that 120=24 ·5, . . ., 995=199·5. Therefore, the number of integers from 120 through 999 divisible by 5 is equivalent to the number of integers from 24 through 199 divisible by 5, or, by the lemma, 199-24 +1 = 176.

Related to this question is an application of the relative frequency interpretation of an event, which we emphasize once more as our justification for studying basic combinatorics. Specifically, we ask, what is the probability of the event a number from 120 through 999 is divisible by 5. According to the relative frequency interpretation of this question, we have

$$\frac{176}{999 - 120 + 1}$$

where the denominator is the obtained by the lemma. Next we turn to the statement and proof of the multiplication rule by beginning with a stating a seemingly unrelated result. Indeed, it is the set theoretic generalization of the law of total probability.

Proposition 8. *Let S be a set and $A_1, A_2, \dots, A_r$ be a partition of length r. Then for any $B \subset S$, we have*

$$|B| = \sum_{i=1}^{r} |B \cap A_i|$$

To prove this proposition, we proceed by the perspective that informs us two sets have the same size if and only if there exists a bijection between them. That, together with the agreement that the cardinality of a union of disjoint sets is equal to the sum of the cardinalities of the sets in the union compels us to consider whether bijection

$$f : B \to \cup_{i=1}^{r} B \cap A_i$$

exists, since the induced sequence of sets $B \cap A_1, B \cap A_2, \dots, B \cap A_r$ are pairwise-disjoint

Let us define f on B by $f(b) = b^i$, where i is the unique index of the partition such that $b \in A_i$. Then with this definition, f is an injection, for if $b^i = b^j$ then $i = j$ by the hypothesis the A_i are a partition of length r, but then $b = b$ by definition of f. Similarly, f is surjective, for suppose $b \in \cup_{i=1}^{r} B \cap A_i$, then there exists a unique i such that $b \in B \cap A_i$, but then $b^i = f(b)$ for some $b \in B$. As b was chosen arbitrarily, f is surjective. Therefore, f is a bijection, as it is both an injection and a surjection, and so, we have $|B| = | \cup_{i=1}^{r} B \cap A_i| = \sum_{i=1}^{r} |B \cap A_i|$, as desired.

An important corollary of this result is when we take the trivial case that $B = S$, for then $|S| = \sum_{i=1}^{r} |A_i|$, as $A_i = S \cap A_i$. Let us add to our repetoire by declaring a partition of length r is *of size n_i* if $|A_i| = n_i$, for $1 \leq i \leq r$. Then we have the following famous result, the so-called addition rule.

Proposition 9. *Let S be a finite set, say $n = |S|$, together with a partition of length r of size n_i. Then*

$$n = \sum_{i=1}^{r} n_i$$

There is nothing to prove as the addition rule is a special case of the proposition. There is also a difference rule, dual to the addition rule. Define $A \setminus B = A \cap B^c$, for some pair $B \subset A \subset S$.

Proposition 10. *With notation as above, $|A \setminus B| = |A| - |B|$*

The proof is again an application of the proposition, this time taking $B, A \setminus B$ as a partition of length two of A itself. The result follows from the addition rule solving for $|A \setminus B|$.

The important application of these results that we have in mind is to stating and proving the multiplication rule, the principle underlying our counting arguments throughout this chapter. To achieve this aim, we shall digress once more, this time to computing the cardinality of the Cartesian product of a sequence of sets. Here we begin to develop a theme that will culminate in our study of Markov chains.

Proposition 11. *Let $A_1, A_2, \dots, A_k$ be a sequence of pairwise-disjoint sets such that $n_i = |A_i|$ and $S = A_1 \times A_2 \times \cdots \times A_k$ their Cartesian product. Then*

$$|S| = \prod_{i=1}^{k} |A_i|$$

where $\prod_{i=1}^{k} n_i = n_1 \cdot n_2 \cdots n_k$ *is the notation for the product of the numbers* n_i.

We will not prove this statement in its entirety, as a full proof would require induction. Informally, however, we shall first proceed to exhibit a bijection f between the sets A_2 and $a \times A_2$, where $a \in A_1$ and with notation as in the proposition otherwise. So, for $f : A_2 \to a \times A_2$ let us define for $b \in A_2$ the function $f(b) = (a, b) \in a \times A_2$. Clearly f is both injective and surjective as $A_1 \cap A_2 = \emptyset$. Therefore, as usual, $|A_2| = |a \times A_2|$.

Second, notice that our above choice of $a \in A_1$ was arbitrary. Moreover, we can decompose $A_1 = \cup_{i=1}^{n_1} a_i$, where $|A_1| = n_1$. As such, the sequence $a_1 \times A_2, a_2 \times A_2, \ldots, a_{n_1} \times A_2$ is a partition of $A_1 \times A_2$ of length n_1. Therefore,

$$|A_1 \times A_2| = \sum_{a_i \in A_1} |a_i \times A_2|$$

Now, the cardinalities are all equal by the transitivity of the relation "there exists a bijection." To wit, $A_2 \cong a_i \times A_2 \cong a_j \times A_2$ for any $a_i, a_j \in A_1$ and where $\cong$ means "there exists a bijection". Since $|A_1| = n_1$ and $|A_2| = n_2$ by hypothesis,we have

$$\begin{aligned} |A_1 \times A_2| =& \sum_{i=1}^{n_1} n_2 \\ =& n_1 \cdot n_2 \end{aligned}$$

as desired.

As mentioned at the outset, one merely requires the induction hypothesis to finish the proof, which would proceed in exactly the same manner up to replacing A_1 by $A_1 \times \cdots A_{n-1}$. We have covered enough material now to state the multiplication rule. However, we no longer must prove it, as we have accomplished this already. Indeed, the proof of the multiplication is merely to forget the labels we added to the sets in the above proposition for its statement. In other words, abstractly, the proposition *is* the multiplication rule. However, such an austere presentation is inappropriate for what we have in mind, so we phrase matters differently in order for the proposition to comport with the counting arguments we have in mind.

Let us *define* a process to be the k-fold Cartesian product $S = A_1 \times A_2 \times \cdots \times A_k$. Furthermore, we assert that the *performance of a process* is, abstractly, a k-tuple, say $(a_1, a_2, \ldots, a_k) \in S$. In particular, one could infer from this definition that a process consists of k mutually independent steps, where a step is taken to be coordinate of a k-tuple. In order to distinguish processes from the abstract rendering of elements of Cartesian products we shall write *words* or concatenations of the elements of the k-tuple, instead. Let us suppose further there are n_i ways to perform the i-th step, for $1 \leq i \leq k$. In terms of S, this means that $|A_i| = n_i$. Then the *multiplication rule* states there are

$$\prod_{i=1}^{k} n_i$$

ways to perform the process. The proof is obviously the same as the above proof giving the cardinality of the Cartesian product S. We have the the following theorem as a corollary.

Theorem 5. *(Multiplication Rule) Let $A_1, A_2, \dots, A_k$ be a sequence of pairwise-disjoint sets up to indexing, such that $n_i = |A_i|$ and $S = A_1 \times A_2 \times \dots \times A_k$ their Cartesian product. Define S to be a process, A_i a step in the process, and a performance of the process to be a word $a_1 \cdot a_2 \cdots a_k$ obtained by concatenation of elements such that $a_i \in A_i$ for each $1 \leq i \leq k$. Then there are $n_1 \cdot n_2 \cdots n_k$ performances of the process S.*

There is nothing to prove for, as mentioned above, the theorem is nothing more than a re-phrasing intended for applications to a statement we have already proven that computes the cardinality of the a Cartesian product of pairwise-disjoint sets. However, the hope is these labels help one to organize their arguments in order to learn the basics of counting. In particular, the phrase "up to labeling" is intended to suggest two sets in this sequence underlying a process are disjoint if their indices are distinct. This will allow us to use recursive arguments below to count various arrangements of the elements of a set, for we are then able to distinguish elements of a performance of a process by the step in which they are introduced or abstractly speaking, their index.

To that end, we shall apply the multiplication rule to count the number of permutations of a set S of size n. A *permutation* is our first example of an *arrangement*, which is a sequence of its elements with a prescribed order and method of selecting elements for inclusion. In the case of a permutation of S, a permutation of its elements is an ordered list of all of its elements selected without replacement from S. As all of the elements are selected for inclusion in the sequence, what distinguishes one such list from another is the order in which they are selected, as duplicate selections are precluded by the prescription that replacement is not allowed.

So to count the number of permutation of a set S or size n, we want to exhibit a permutation as a performance of a process, which according to the multiplication rule, is defined as a word with respect a Cartesian product of pairwise-disjoint sets up to indexing. Accordingly, impose an order on the set S simply by labeling the finite number of its elements, say n, by natural numbers. This order will determine the steps or indices of the process. To wit, write $S = \{a_1, a_2, \dots, a_n\}$. Next, let us define a sequence of pairwise-disjoint sets up to indexing by the following recursion rule. Define $A_1 = S, A_2 = S(a_1), \dots, A_i = S(a_1, a_2, \dots, a_{i-1}), \dots, A_n = S(a_1, a_2, \dots, a_{n-1}) = a_n$, where the notation $S(a_1, a_2, \dots, a_{i-1})$ is defined to be $S \setminus \{a_1, a_2, \dots, a_{i-1}\}$ to indicate the preceding $i-1$ elements of S where selected, without replacement. Then with notation as above, a permutation is a word with respect to Cartesian product

$$\bigtimes_{i=1}^{n} A_i$$

By construction, $|A_i| = n - i + 1$, so that, by the multiplication rule, the number of permutations of the set S is

$$n! = \prod_{i=1}^{n} n - i + 1$$

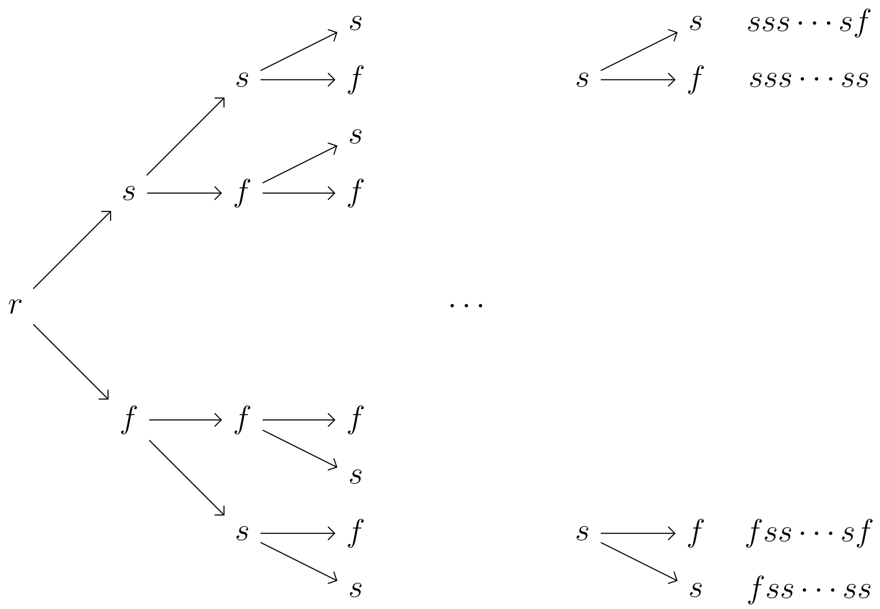
Expanding this product gives the famous $n!$ or factorial symbol, interpreted combinatorially as the number of permutations of a set of size $n \leq 0$. In particular, as $|\emptyset| = 0$ we have, vacuously, that 0!=1, for there is one arrangement of no elements. In subsequent sections we shall apply the multiplication rule less formally for the sake of clarity. Nevertheless, in what follows, the reader may reconstruct a more formal argument in terms of the aforementioned definition of process.

An illustration of a less formal application is to counting the number of subsets of a set S of size n. Indeed, the process of forming a subset is determined by deciding whether to include a particular element. To simplify the demonstration, impart an order to the elements of S so that one may visualize a subset as an n-tuple $(a_1, a_2, \ldots, a_n)$ such that either a_i is included in this n-tuple or it is not. If it is not, then leave the i-th position blank. Accordingly, there are n steps in this process, and two ways to perform each step. Therefore, by the multiplication rule, we have that the number of subset of S is 2^n. This figure is more than just a conclusion for us for it is also the cardinality of the *power set*. Indeed, we have proven that $|2^S| = 2^n$.

We end this section by indicating how a process may be visualized informally as a tree diagram. We stress the informality of this presentation according to the rigorous or formal alternative, which requires initially the definition of a graph. We shall provide this material later in chapter 6 yet it is easy enough to work with these concepts now, as long as one agrees to ignore ambiguities and the imprecision of the informal treatment. Granting this approach, let us consider a process as the set of branches in a tree diagram. Indeed, such a tree diagram is a special case of the more general notion of *tree* introduced later, and for this reason, we will call it a *process tree*.

A process tree is a visual illustration of the elements of a Cartesian product of sets. The branches of the tree correspond to tuples or elements of the Cartesian product. The set of elements contained in a factor of such a product, $S = \bigtimes_{i=1}^{k} A_i$, say A_i, appear written vertically at the i-th level, where level is counted from left to right starting at $i = 1$ and ending at $i = k$. One writes the elements that comprise the vertices of a particular branch to the right of its terminal vertex in a column down the right-hand side of the process tree. Such words furnish the correspondence between elements of the Cartesian product and branches of the process tree.

Consider the following illustration of the process tree associated to the Cartesian product $\bigtimes_{i=1}^{n} B_i$ of n sets $B_i = \{s, f\}$ for all $i = 1, 2, \ldots, n$



where for purposes of illustration we indicate the 0-th level of this process tree with an r so that the branches all descend from a common source. Such process trees shall feature prominently below when we consider Bernoulli processes and our main examples of parametric families of discrete random variables.

2.1 WORKED EXAMPLES

1. 2.1 Suppose a computer installation has four input/output units, say a, b, c, and d and three central processing units, say x, y, and z.

   (a) How many ways are there to pair an input/oupt unit with a central processing unit?

   (b) List three performances of this process.

   Solution (a) To answer this question, we must determine the number of steps in this process and the number of ways to perform each step, then, by the multiplication rule, we take the product of these figures to find the total number of ways to pair input/output units with central processing units. In more sophisticated terms, we must determine the factors of the Cartesian product product corresponding to this process. Namely, $A_1 = \{a, b, c, d\}$ and $A_2 = \{x, y, z\}$ so that the entire process is represented by the Cartesian product $S = A_1 \times A_2$ whose cardinality $|S| = |A_1| \cdot |A_2| = 12$ is the total number of ways to pair input/output units with central processing units.

   Solution (b) Proceeding by definition, which is to identify a perforance of a process with a tuple in the Cartesian product representing the same, we must simply write down three 2-tuples in S. So, for example, $(a, x), (a, y)$, and (a, z) would do as examples of performances of the process. However, so would $(b, x), (b, y)$, and (b, z).

2. 2.1 Conaider the folliwing nest loop which runs for $i = 1$ to $i = 4$ and for $j = 1$ to $j = 3$, next j, next i.

   (a) How many total iterations of this nested loop exist?

   Solution (a) Perhaps this solution is best illustrated by a process tree, but nevertheless, the outer loop corresponding to i is iterated four times and the inner loop corresponding to j is iterated times, once each time the outer loop is iterated. Therefore, by the multiplication rule, there are two steps, the first performed four times, the second performed three times. Therefore, there are twelve total iterations of the nested loop.

3. 2.1 A combination lock has integer labels around a wheel consisting of the integers from 0 through 39, called its dial. A combination code consists of a 3-tuple of integers selected from 0 through 39, in order, with repetition allowed.

   (a) How many total combination codes exist for a combination lock?

   Solution (a) We proceed by the multiplication rule to count the total number of combination codes by first identifying that there are three steps in this process, one for each selection of an integer from 0 through 39. The number of ways to perform each step is equivalent to the number of integers from 0 through 39, according to the fact that a performances amounts to selecting an integer. There are 39-0+1=40 ways to do this. Therefore, by the multiplication rule, there are 40^3 total combination codes.

2.2 The Definition of an Arrangement and r-Orderings

In this section we introduce the first of four arrangements of a set finite S. To this end, let us first define what an arrangement is.

Definition 19. *Let S be a finite set, then an arrangement is a sequence of its elements such that*

1. *elements in the sequence are either ordered or unordered*
2. *elements in the sequence are selected from S either with replacement after selection or without replacement after selection for inclusion in the sequence*

Plainly the process of forming arrangements of S can be done in four ways since in the first step we must choose whether the sequence is ordered and in the second step we choose whether or not to replace elements selected for inclusion. This argument shall prefigure the remainder of this chapter, as we characterize each of these four arrangements and consider topics related to them. The first of these four arrangements that we shall consider is called an *r-ordering.*

Definition 20. *Let S be a finite set and $n = |S|$. We say an r-ordering of S is an ordered arrangement of r of its elements with replacement after selection for inclusion.*

Immediately one should notice that the integers r and n are independent of each other because the selection of elements for inclusion in an r-ordering from S is inexhaustible; that is, one selects elements for inclusion with replacement. Next we count the number of r-orderings of a finite set S. We proceed by the multiplication rule.

Theorem 6. *Let S be a finite set and $n = |S|$. The number of r-orderings of S is n^r.*

The proof is a straightforward application of the multiplication rule once we identify the process of forming an r-ordering of S. Indeed, a step in this process corresponds to the selection of an element. Therefore, there are r steps in this process. Further, there is n ways to perform each step by the hypothesis that we are selecting elements for inclusion with replacement of the same. Therefore, the theorem follows by simplifying the r-fold product of n with itself obtained from the multiplication rule.

2.2 WORKED EXAMPLES

1. 2.2 An easy example is to count the number of 6-orderings of the set $S = \{a, b, c, d, e\}$.

   (a) Count the number of 6-orderings of the set $S = \{a, b, c, d, e\}$.

   Solution (a) By the theorem, since $n = 5$, we have there are $5^6 = 15625$.

2. 2.2 Let AD be the set of Roman alphabet letters and the numerals 0 through 9. An automobile license plate is a 7-ordering of the set AD.

   (a) Count the number of license plates.

   Solution (a) Again, we proceed by the theorem to obtain that there are 36^7 license plates, since there are $|AD| = 36$ elements in AD.

3. 2.2 Let $f : S \to T$ be a set function and $|S| = n, |T| = m$.

   (a) Count the number of such functions f.

Solution (a) Our objective is to count the number of such f as an r-ordering. More specifically, consider the process of forming a set function. By definition, a set function is a well-defined rule that assigns elements of S to elements of T. As the rule is well-defined, we are limited to considering only rules that uniquely assign each $a_i \in S$ to an element $b_j \in T$. Accordingly, the steps of this process are the assignment of elements in S to elements in T and the number of ways each $a_i \in S$ may be assigned to an element in T is self-evidently m, that is, the cardinality of T. Therefore, the number of set functions from S to T is equivalent to the number of n-orderings of T, that is, m^n.

Notice that we placed no restrictions on these functions beside their definition, that is, that they are well-defined. We shall use this argument below to show that the number of bijections of a set with itself is $n!$ and to prove thereby, in a combinatorial manner, that $n! \leq n^n$, as there are clearly more set functions from a set to itself than bijections of the same.

2.3 r-Permutations

In this section we shall consider the following kind of arrangement, which is a generalization of the permutation discussed in section 2.2.

Definition 21. *Let S be a finite set and $n = |S|$. We say an r-permutation of S is an ordered arrangement of r of its elements without replacement after selection for inclusion.*

This arrangement is the second of the four arrangements that structure our approach to combinatorics in this chapter. Notice, as the second in our sequence, the alteration in hypotheses implies that $r \leq n$, for if $r = n$ the elements available for selection in the arrangement have been exhausted.

Counting the number of permuations of a finite set S generalizes the argument in section 2.2. Indeed, recall the notation from that section and define $A_1 = S, A_2 = S(a_1), \ldots, A_r = S(a_1, a_2, \ldots, a_{r-1})$, where the notation $A_i = S(a_1, a_2, \ldots, a_{i-1})$ is defined to be $S \backslash \{a_1, a_2, \ldots, a_{i-1}\}$ to indicate the preceding $i-1$ elements of S where selected, without replacement. Then with notation as above, an r-permutation is a word with respect to Cartesian product

$$\bigtimes_{i=r}^{n} A_i$$

By construction, $|A_i| = n - i + 1$, so that, by the multiplication rule, the number of r-permutations of the set S is

$$n \cdot (n-1) \cdots (n-r+1)$$

since there are $(n-r+1)$ integers from r through n.

While this formula is the correct conclusion from the multiplication rule, it is not convenient for applications or memorization, for that matter. To that end, we further simplify it as follows

$$n \cdot (n-1) \cdots (n-r+1) \frac{(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!}$$

since $n \cdot (n-1) \cdots (n-r+1) \cdot (n-r)! = n!$ by definition. We have now proven the following theorem.

Theorem 7. *Let S be a finite set and $n = |S|$ and a non-negative integer $r \leq n$. The number of r-permutations of S is*

$$nPr = \frac{n!}{(n-r)!}$$

We can apply this formula in some easy examples. For instance, let us count the number of 3 permutations of the set $S = \{a, b, c, d, e\}$. As $|S| = 5$ we have, by the theorem, that there are $\frac{5!}{2!} = 5 \cdot 4 \cdot 3 = 60$. Moreover, suppose that one of the letters of such a 5-permutation of S must be b, then how many such 5-permutations are there of S? As one expects, since one of the positions in the arrangement is fixed, such arrangements are tantamount to 4-permutations of the set $S' = \{a, c, d, e\}$ so that, by the theorem, there are $\frac{4!}{2!} = 12$.

Perhaps an advantage of the theorem is that it allows one to treat combinatorial problems algebraically. To that end, we can show that

$$\frac{n!}{(n-2)!} + \frac{n!}{(n-1)!} = n^2$$

according to the follwing simplification argument.

$$\begin{aligned} \frac{n!}{(n-2)!} + \frac{n!}{(n-1)!} &= \frac{n!(n-1) + n!}{(n-1)!} \\ &= \frac{n! \cdot n}{(n-1)!} \\ &= n^2 \end{aligned}$$

Last, let us consider the number of different ways in which five ranked prize winners may be chosen from a group of one hundred people. Of course, such a selection of ranked winners is ordered by the ranking. Therefore, such a selection is a 5-permuation of the 100 people. Accordingly, there are $\frac{100!}{95!} = 9,034,502,400$. Hopefully, the great magnitude of this answer convinces the reader that sometimes such large computations are better left in their original notation.

2.3 WORKED EXAMPLES

1. 2.3 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

2. 2.3 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3. 2.3 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2.4 r-Combinations

In this section we both introduce and count the number of so-called r-combinations of a finite set. This particular arrangement is important, for it underlies the binomial theorem and is also used to compute its several variable generalization. Both theorems have a role to play in parametric statistics, but the greatest is perhaps played by the binomial theorem. Indeed, our proof in chapter 5 that the binomial random variable satisfies the induced Kolmogorov axioms depends upon its formula. Given the major role of the binomial random variable in modern mathematics, the importance of combinations cannot be overstated. Therefore let us consider the following definition.

Definition 22. *Let S be a finite set and $n = |S|$. We say an r-combination of S is an unordered arrangement of r of its elements without replacement after selection for inclusion.*

The process of forming an r-combination of S consists of but two steps. Observe, first we must choose r elements of S. Choosing r elements requires no order, so in the second step, let us choose their order. As we know, there are $r!$ ways to put r elements in order. Accordingly, if nCr is the number of ways to choose r elements of S, we have by the multiplication rule that

$$nCr \cdot r! = nPr$$

but then nCr is precisely the figure we meant to count. This gives the following theorem.

Theorem 8. *Let S be a finite set, $n = |S|$, and r a non-negative integer $r \leq n$. The number of r-combinations of S is*

$$nCr = \frac{n!}{r!(n-r)!}$$

Of course the theorem follows from solving for nCr in our application of the multiplication rule to the task of counting the number of unordered arrangements of S without replacement of selection. Another famous notation for nCr is $\binom{n}{r}$ and both symbols are pronounced as "n choose r." We shall reserve the latter notation for the binomial theorem.

Let us consider several examples of r-combinations. For instance, the number of ten person committees of United States senators would be an example of 10-combinations of the United States senate, which consists of 100 members. Accordingly, there are $100C10 = \dfrac{100!}{10!90!} = 17,310,309,456,440$. Perhaps this figure sheds some light on the reason why politics rather than arbitrary selection are required to narrow down the list of possible committee assignments!

Next, let us consider forming a team of five from a group of twelve colleagues. Suppose two members of this group insist on being together on any five person team for which either is selected. How many five person teams can be formed from this group?

The solution to this question is not as straightforward as simply computing a single combination of the group of twelve colleagues. Indeed, we may partition the set of five person teams by those that include the pair of colleagues and those teams which do not. The number of the former five person teams is a 3-combination of 10 of the colleagues, for the pair is not included. The number of the latter five person is a 5-combination of the 10 colleagues, because the pair is not included. As such, there are $10C3 + 10C5 = \frac{10!}{3!7!} + \frac{10!}{5!5!} = 120 + 225 = 372$ five person teams formed from the twelve colleagues.

Let us suppose a similar group of twelve people is partitioned according the color of their shirts; five people are wearing blue shirts and seven people are wearing pink shirts. How many five person teams can be formed from this group that contain at least one blue shirt?

To solve this problem, we can proceed by the difference rule. Let us write S for the set of all five person teams formed from this group, and A for the set of five person teams with no members wearing a blue shirt. Then, logically speaking, the set of teams with at least one blue shirt is characterized by the complement of A in S, namely, by the set $S \setminus A$. Moreover, tautologically, A and its complement in S partition S, we have, by the difference rule, that $|S \setminus A| = |S| - |A|$. Now as the cardinality of the set of fiver person teams without prescription are the 5-combinations of the twelve person group, we have $|S| = 12C5$. Further, as the number of five person teams with no blue shirts is the number of 5-combinations of remaining seven people, so $|A| = 7C5$. Therefore $|S \setminus A| = 12C5 - 7C5 = 771$.

Last, let us consider counting various hands in the game of poker. In particular, proceeding by the relative frequency interpretation of probability, we shall compute the probabilities of the hands we present in our examples. Recall that the game of poker is a card game that consists of a deck of 52 playing cards, partitioned into four suites, called hearts, diamonds, clubs, and spades, whose cards are labeled $2, \ldots, 10$ and J, Q, K, A. A *hand* is a subset of size five cards taken from the deck. In other words, a hand is a 5-combination of the deck. Hence, the total number of hands in a game of poker is $52C5$-we will require this fact in a moment.

Given our description of the game, let us both count the number of two pair hands, defined to be hands that consist of two pairs, where a pair of cards is defined to be two cards whose labeling matches, and a fifth card whose label is different from that of either of the pairs. Let us proceed by the multiplication rule. One should notice there are several distinct ways to apply it to this problem, but each application arrives at the same result, so by the fundamental theorem of counting they are equivalent. In any case, we should divide this process into four steps as follows.

First, let us choose the labels of both pairs. As there are thirteen labels, there are $13C2$ ways to perform this step. Second, choose the suites of the smaller pair-traditionally the order of the labels is the one indicated above. Since suites partition the deck, and there are two cards in a pair, we can perform this step in $4C2$ ways. Third, choose the suites for the larger pair. Again, there is $4C2$ ways to perform this step. Fourth, choose the label of the fifth card. Since it must be distinct from that of the paired labels, there are $11C1$ ways to perform this step. Last, choose the fifth card's suite. Clearly there is $4C1$ ways to perform this step. Therefore, by the multiplication rule, there are $13C2 \cdot 4C2^2 \cdot 11C1 \cdot 4C1 = 123,552$ two pairs hands in poker. Moreover, the probability is obtained by the relative frequency of this event, so that probability of the event "a two pair hand" is obtained by the quotient $\frac{123,552}{52C5}$ or approximately 4.7 %.

2.4 WORKED EXAMPLES

1. 2.4 Determine the probability the following events:

(a) $A =$

(b) $B =$

(c) $C =$

We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

2. 2.4 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3. 2.4 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2.5 The Binomial Theorem

In this section we state and prove the famous binomial theorem. As mentioned above, this formula underlies the proof that the probability space induced by the binomial random variable satisfies Kolmogorov's axioms. This result alone is enough in the author's opinion to justify its inclusion. Beside its renown, one of the proofs shall exhibit the technique of induction. Our first proof shall utilize the combinatorial interpretation of the number of r-combinations of a finite set.

Indeed, let us consider a set S of size n together with a labeling of its elements, say $S = \{a_1, a_2, \ldots, a_n\}$. Plainly an arbitrary subset $A_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,r}\} \subset S$ is an r-combination of the elements of S, as one selects elements without replacement for inclusion, and there is no particular order in which they must be arranged. Therefore the combinatorial interpretation of nCr is that it counts the number of subsets of S of size r.

As an illustration of this perspective's application to the binomial theorem, we shall prove the symmetry of the *binomial coefficient* with respect to its lower argument. Namely, in the context of counting problems germane to the number of subsets of various sizes, write

$$nCr = \binom{n}{r}$$

For a set S of size n, let $2^S(i)$ be the set of subsets of S of size i.

Lemma 2.1. *With notation as above,* $|2^S(r)| = |2^S(n-r)|$. *In particular,*

$$\binom{n}{r} = \binom{n}{n-r}$$

As usual, to prove two sets have the same cardinality we shall exhibit a the existence of a bijection between the same. To this end, define $f : 2^S(r) \to 2^S(n-r)$ by $f(A_i) = S(A_i) = S \setminus \{a_{i,1}, a_{i,2}, \ldots, a_{i,r}\}$. The map is both injective and surjective by the law of the excluded middle-namely, an element is either itself or it is not and never both-together with the definition of set complement. Therfore, $\binom{n}{r} = \binom{n}{n-r}$, as desired. Next, we prove Pascal's identity, which is the result that validates the inductive step in the proof of the binomial theorem.

Theorem 9. *Let S be a finite set, $n = |S|$, and r a non-negative integer $r \leq n$. Then*

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$$

To prove this identity, we shall adopt the notation of the previous paragraphs in this section. Let $a_i \in S$ and $2^S(r) \setminus \{a_i\}$ for the set of subsets of size r that do not contain a_i and $2^S(r - a_i)$ for the set of subsets of size r that do contain a_i. Plainly these sets partition $2^S(r)$, so by the addition rule, we have

$$|2^S(r)| = |2^S(r) \setminus \{a_i\}| + |2^S(r - a_i)|$$

Observe that $2^S(r) \setminus \{a_i\}$ consists of r-combinations of a set of size $n-1$ for a_i is excldued. Therefore, $|2^S(r) \setminus \{a_i\}|$ is $\binom{n-1}{r}$. Similarly, $2^S(r - a_i)$ consists of $r-1$ combinations of a set of size $n-1$ as a_i *must* be included. Therefore, $|2^S(r - a_i)|$ is $\binom{n-1}{r-1}$. Thus, by the addition rule, we have

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$$

as desired.

We are now in a position to prove the binomial theorem. We shall prove it twice: once by induction and a second time combinatorially. The first proof requires establishing the basis for induction, which we hope sheds light upon the combinatorial argument. The induction step requires Pascal's identity, which is why we begin with it.

Before we proceed to prove the theorem, let us recall the technique of *proof by induction.* Let $\pi(n)$ be any logical statement that depends on an integer $n \geq n_0$, where n_0 is some fixed integer. The $\pi(n)$ is true for all such n provided both $\pi(n_0)$ and if $\pi(n)$ is true, then $\pi(n+1)$ for all $n \geq n_0$. The statement $\pi(n_0)$ is often referred to as the basis for induction and $\pi(n)$ the induction hypothesis, for one is allowed to assert this statement in order to demonstrate $\pi(n+1)$. There are several equivalent formulations of this technique in the literature, but this one shall suffice for our purposes. Let us now exhibit this technique to prove the following theorem.

Theorem 10. *Let $n \in \mathbb{N}$, then for any binomial $(x+y)$, we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

Let $\pi(n)$ be the statement of the theorem. We shall establish the basis for induction by choosing $n_0 = 4$, although we could choose other more or less convenient values, since it is true for any $n \geq 0$ as above. So, consider the expansion

$$\begin{aligned}
(x+y)^4 =&(x+y)(x+y)(x+y)(x+y)\\
=&x^4+\\
&x^3y + x^2yx + xyx^2 + yx^3+\\
&x^2y^2 + x^2y^2 + xy^2x + yx^2y + y^2x^2 + y^2x^2+\\
&y^3x + y^2xy + yxy^2 + xy^3+\\
&y^4\\
=&\binom{4}{0}x^4 + \binom{4}{1}x^3y + \binom{4}{2}x^2y^2 + \binom{4}{3}xy^3 + \binom{4}{4}y^4\\
=&\sum_{k=0}^{4}\binom{4}{k}x^{4-k}y^k
\end{aligned}$$

Notice several lines above are apparently needless, but they shall help clarify our combinatorial proof below. Nonetheless, at present, let now assume that $\pi(n)$ is true and show that it implies $\pi(n+1)$. We shall use Pascal's identity, as follows.

$$\begin{aligned}(x+y)^{n+1} =&(x+y)^n(x+y)\\ =&\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k(x+y)\\ =&\sum_{k=0}^{n}\binom{n}{k}x^{n-k+1}y^k+\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^{k+1}\\ =&x^{n+1}+\sum_{k=1}^{n}\binom{n}{k}x^{n-k+1}y^k+\sum_{k=1}^{n}\binom{n}{k-1}x^{n-k+1}y^k+y^{n+1}\\ =&x^{n+1}+\sum_{k=1}^{n}\left(\binom{n}{k}+\binom{n}{k-1}\right)x^{n+1-k}y^k+y^{n+1}\\ =&x^{n+1}+\sum_{k=1}^{n}\binom{n+1}{k}x^{(n+1)-k}y^k+y^{n+1}\\ =&\sum_{k=0}^{n+1}\binom{(n+1)}{k}x^{(n+1)-k}y^{k+1}\end{aligned}$$

which is indeed $\pi(n+1)$, as desired. Notice the penultimate line required Pascal's identity.

The combinatorial proof is suggested by the basis for induction, as indicted earlier. Indeed, $(x+y)^n=(x+y)\cdots(x+y)$ n-times. Each monomial collects one variable from each binomial term in the product, therefore, each monomial is of degree n. In particular, it is of the form $x^{n-k}y^k$ for some $k\in\{0,1,\dots,n\}$ representing the number of times y was chosen. We see this process for determining monomials in the basis for induction when $n=4$. Continuing then, the number of such monomials is equivalent to the number of subsets of n variables of size k, namely $\binom{n}{k}$. Accordingly, combining like terms gives

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k$$

as desired.

Let us now establish a few identities and present a few examples.

2.5 WORKED EXAMPLES

1. 2.5 Determine the probability the following events:

   (a) $A=$

   (b) $B=$

   (c) $C=$

   We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

2. 2.5 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3. 2.5 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2.6 The Multinomial Theorem and Ordered Partitions

In this section, we mean to generalize the binomial theorem by considering a new type of arrangement, predicated upon the previous arrangements, and count the number of those. The generalization we seek is that of the *multinomial theorem.*

Definition 23. *Let S be a finite set, $n = |S|$, and $\{A_i\}_{i=1}^{r}$ a partition of S of length r. We say such a partition is of size n_i if each cell $|A_i| = n_i$. We say a permutation of S such that the first n_1 elements belong to A_1, the second n_2 elements belong to A_2 and so and so forth until the last n_r elements belong to A_r is an ordered partition of S.*

An alternative, although perhaps more opaque, definition of an ordered partition of S is a bijection of S with itself, where a label is added to indicate how an element of S is assigned to a cell in the given partition of length r and size n_i. Let us now count the number of ordered partitions of S.

This process consists of r steps, one for each cell of the partition of S. Moreover, assigning n_i elements of S without replacement to a particular cell imposes the following counting argument. In the first step, choose a subset of size n_1. There are $\binom{n}{n_1}$ ways to perform this step. In the second step, choose a subset of size n_2 from the remaining $n - n_1$ elements, for selection is made withou replacement in a permutation. There are $\binom{n-n_1}{n_2}$ ways to perform this step. And so on and so forth until the final or r-th step in this process. In this step, choose a subset of size n_r from the remaining $n - n_1 - n_2 - \cdots - n_{r-1}$ elements. Altogether, by the multiplication, there are

$$\tbinom{n}{n_1} \cdot \tbinom{n-n_1}{n_2} \cdots \tbinom{n-n_1-n_2-\cdots-n_{r-1}}{n_r}$$

ways to perform this process of forming ordered partitions of S. Simplifying this argument, we have, in terms of the binomial coefficients,

$$\frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdots \frac{(n-n_1-n_2-\cdots-n_{r-1})!}{n_r!(n-n_1-n_2-\cdots-n_{r-1}-n_r)!}$$

which cancels pairwise, denominator against subsequent numerator, except in the r-th factor. There, in the denominator, we notice that, by the addition rule, $n-n_1-n_2-\cdots-n_{r-1}-n_r = 0$ since $n = n_1 + n_2 + \cdots + n_r$ by the hypothesis that $\{A_i\}$ is a partition of length r and size n_i. Altogether, we have now prove the the following theorem.

Theorem 11. *Let S be a finite set, $n = |S|$, and $\{A_i\}_{i=1}^{r}$ a partition of length r and size n_i. Then the number of ordered partitions of S is given by*

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdots n_r!}$$

We refer to the figure in the theorem, obtained by the simplification of our multiplication rule argument, as the *multinomial coefficient.* The reason for this name is that the number of ordered partitions determines the coefficients in the expansion of the n-th power of r-fold multinomial.

Theorem 12. *Consider the multinomial $x_1 + x_2 + \cdots + x_r$, then the expansion of its n-th power is given by the formula*

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{(n_1,n_2,\ldots,n_r)\in\mathbb{Z}^r | n_1+n_2+\cdots+n_r=n} \binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdots n_r!} x_1^{n_1} \cdot x_2^{n_2} \cdots x_r^{n_r}$$

An interesting but trivial computation shows that, by setting $x_i = 1$, we have that

$$r^n = \sum_{(n_1,n_2,\ldots,n_r)\in\mathbb{Z}^r | n_1+n_2+\cdots+n_r=n} \binom{n}{n_1, n_2, \ldots, n_r}$$

2.6 WORKED EXAMPLES

1. 2.6 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2. 2.6 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

3. 2.6 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2.7 r-Unorderings

The last type of arrangement we consider is a so-called r-unordering. These are arrangements are perhaps initially the most counter-intuitive when introduced, but nevertheless, commonly occur in recognizable counting problems.

Definition 24. *Let S be a finite set and $n = |S|$. We say an r-unordering of S is an unordered arrangement of its elements with replacement after selection for inclusion.*

The method of counting such arrangements is of interest in its own right, as the method is a common and effective one of combinatorics. It is the so-called *stars and bars* argument. We shall characterize an r-unordering as a combination of stars and bars and count the number of those instead. Indeed, let $S = \{a_1, a_2, \ldots, a_n\}$ be the elements of S. We can represent these elements by the compartments induced by $n-1$ bars. Indeed, to see this, remove the a_i notation from the sequence

$$a_1|a_2|\cdots|a_n$$

The efficacy of this translation is that replacement after selection for inclusion is easy to represent. To do this, we insert a star in the compartment corresponding to $a_i \in S$ each time it is selected for inclusion. For example, an r-unordering where a_1 is selected twice, a_2 is selected three times, etc. a_{n-1} is selected once and a_n is selected not at all could be represented by the following sequence of stars and bars, viz.

$$\star\star\,|\,\star\star\star\,|\cdots|\,\star\,|$$

Here we have informally established a bijection between the set of r-unorderings of S and sequences of $n-1$ stars and r bars. We may therefore count the number of the latter kind of sequences to count the former number of arrangements.

Observe that a sequence of $n-1$ bars and r stars corresponding to an r-unordering of S is a *combination* of the set of $n-1$ bars and r stars. In particular, as there are r elements in an r-unodering of S, it is an r-combination of the set of $n-1$ bars and r stars. Therefore, as r-unorderings are in bijective correspondence with sequences of $n-1$ bars and r stars, there are

$$\binom{n-1+r}{r}$$

r-unorderings of S. Notice the formula for the number of such arrangements reflects the fact that r is independent of n, as replacement of selection for inclusion is allowed. Indeed, since the lower index of the binomial coefficient depends on r alone, r may be an arbitrary non-negative integer for n fixed. Let us now complete this chapter with some examples of r-unorderings.

Consider the host of a party who wishes to set out fifteen assorted cans of soft drinks for his guests. he stops at a store that sells five different brands of soft drink. How many different arrangements of fifteen assorted cans of soft drink may the party host set out? The solution to this question is obtained by recognizing an assortment of soft drinks as an unordering of the available brands at the store. Indeed, the host is considering 15-orderings of the 5 brands. As such, there are $\binom{5-1+15}{15} = 3,876$ such assortments for the host to select.

Suppose next that the same host still wishes to present an assortment of fifteen soft drinks, but six cans must be of the root beer brand-how many such assortments are possible, now? What changes is the number of cans we are selecting for our assortment, since 6 are now chosen *ab initio*. So, there are $\binom{5-1+9}{9} = 715$ such assortments in this second case.

Next, let us consider counting non-increasing sequences of integers. To this end, let n be a non-negative integer and consider triples of integers (i, j, k) such that $1 \leq i \leq j \leq k \leq n$, that is, *non-increasing sequences*. As usual our goal is to count the number of such non-increasing sequences. To do this, we must recognize how they are unorderings. Represent such a triple as a combination of bars and stars by two bars to represent the first, second, and third entries in the sequence by compartments and n stars, where one places i stars in the first compartment, j stars in the second, and k stars in the third. It then follows such non-increasing sequences may be represented by 3-combinations of $n-1+3$ stars and bars, or $\binom{n-1+3}{3} = \binom{n+2}{3}$.

Last, let us consider counting the number of integral solutions to linear equations. In particular, let us count the number of such solutions to the equation

$$x_1 + x_2 + x_3 + x_4 = 10$$

as unorderings. We shall proceed as above to count the number of equivalent combinations of stars and bars.

Recall each non-negative integer n is an n-fold sum of the unit 1. We may therefore represent a integral solution to the given equation as a combination of ten stars and three bars. Each compartment demarcated by the three bars represents a variable in the equation. A star in each compartment represents an addition of 1 to obtain the integer substituted for the corresponding variable. For example, $\star\star \mid \star\star\star \mid \star\star\star \mid \star\star$ represents the solution $x_1 = 2, x_2 = 3, x_3 = 3, x_4 = 2$. As such, the number of integral solutions to the given equation is equal to the number of 10-unorderings of 4 variables, which there are $\binom{4-1+10}{10} = 286$ in total. Furthermore, supposing we insist one only consider positive integral solutions, or $x_i \geq 1$, then there are $\binom{4-1+6}{6} = 84$, as four ones are immediately placed in each of the compartments following the previous argument.

2.7 WORKED EXAMPLES

1. 2.7 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2. 2.7 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3. 2.7 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

2.8 Chapter 2 Exercises

1. equivalence class of permutations of the elements in S, where two permutations are equivalent if they produce the same ordered list.

remark: I am bothering to define a "word" as a synonym for "permutation" in our notes to help clarify some of the verbiage, or way things are written, in Gordon's text book. A point of emphasis so far in our lectures has been to identity specific arrangements of elements of a set, so qualifications like whether they are ordered or repetition is allowed are important pieces of information to help solve problems. I feel Gordon's text does not adequately convey these qualifications, therefore the definition is meant to help read both the textbook and books alike it. The phrase "equivalence class" is meant to explain how to identify two permutations which produce the same list. If there are redundant elements in S, then the strict definition of permutation would distinguish ordered lists of elements if the same element with different subscripts was ordered differently in the list. However, by taking equivalence classes, we identify such distinct lists *qua* permutations in the strict sense. In this manner, one is *not* computing simply $|S|!$ to answer these questions. Instead, one needs to use our work that computes the number of allocations of a set S to a partition, or more numerically, the so-called multinomial coefficient formula.

a.) How many words on the set $S = \{\mathbf{F},\mathbf{L},\mathbf{U},\mathbf{F},\mathbf{F}\}$ exist?

b.) How many words on the set $S = \{\mathbf{R,O,T,O,R}\}$ exist if **T** is in the middle?

2. How many numbers can be made each using all the numerals in the set $S = \{1, 2, 2, 3, 3, 3, 0\}$?

3. Five persons, A,B,C,D, and E, are going to speak at a meeting.

   a.) In how many orders can they take their turns speaking if B must speak (sometime) after A?

   b.) In how many orders can they take their turns speaking if B must speak immediately after A?

4. At a table in a restaurant, six people ordered roast beef, three ordered turkey, two ordered pork chops, and one ordered flounder fish. Of course, no two portions of any of these items are absolutely identical. The 12 servings are brought from the kitchen. In how many ways can they be distributed so that everyone gets their correct order?

5. a.) In how many ways can eight people sit at a lunch counter with eight stools?

   b.) In how many ways can four couples sit at the lunch counter if each sits next to one another?

   c.) In how many ways can eight people sit at a round table?

   remark: The idea of a lunch counter is that it is a single row of stools against a counter. So the stools are arranged in a straight line. The intermediate stools have a neighbor on both the left and the right. The stools at the ends have either no left neighbor or no right neighbor.

6. How many non-negative integer solutions are there to the equation

$$x_1 + x_2 + x_3 + x_4 = 30$$

7. A camera shop stocks eight different types of batteries.
   a.) How many ways can a total inventory of 30 batteries be distributed among the eight different types of batteries?
   b.) Assuming that one of the types of batteries is A76, how many ways can a total inventory of 30 batteries be distributed among the eight different types if the inventory must include at least four A76 batteries?

8. Prove that $\binom{n}{r+1} = \dfrac{n-r}{r+1}\binom{n}{r}$ by Computing the symbol on the left hand side to be equal to the symbol on the right hand side of the equality.

HINT: Compute the left hand side until you get the right hand side with $\binom{n}{n-r}$ instead, then use the lecture notes on the binomial theorem section to finish the problem. At least that's how I did it.

9. a.) Completely simplify $\displaystyle\sum_{i=0}^{10} \binom{10}{i} 2^i$

   b.) Expand $(1-x)^6$ and simplify.

   c.) Expand $(x+x^{-1})^5$ and simplify.

3 Conditional Probability

In this chapter we introduce the conditional probability of an event A in an experiment given an event B has already occurred. One could say this introduces a relative notion of the probability of A given B has occurred. Considering the myriad of ways in which such probabilities occur naturally, it is of no wonder that we should desire a precise formulation for ourselves. In terms of chapter one, this chapter generalizes its material by replacing Ω by B with respect to the computation of $P(A)$. Indeed, rather tautologically, one could think of $P(A)$ as the probability of A given that Ω has occurred. Hence, throughout this chapter, we assume the existence of a discrete probability space (Ω, β, P) where β and P have their canonical definitions, in the background. Moreover, we shall refer to this data more informally by referring freely to Ω as an experiment as a shortening of the phrase "discrete probability space." Beside the intrinsic interest of such generalities, a fundamental application is to the definition of stochastic processes and their resolutions into Markov chains in chapter 6, as well as the most famous result of Bayes.

3.1 Independent Events

In this section we introduce sequences of events whose joint probability is a special case of interest in light of the joint probability assigned in general to an arbitrary sequence sequence of events. Whereas the joint probability $P(\cap_i A_i)$ is only formally defined as an application of P to $\cap i A_i \in \beta$, the sequences of events $A_1, A_2, \ldots \in \beta$ we study in this section simplifies this figure both significantly and meaningfully.

Definition 25. *Let $A, B \in \beta$ be events in an experiment Ω, then we say A and B are independent events if*

$$P(A \cap B) = P(A)P(B)$$

One may perhaps recognize that independence is to intersection of events as mutual exclusivity or disjointedness is the union of events. Indeed, by conditions describe a pair of events such that P preserves their operation arithmetically. However, one must not allow this analogy to confuse the two concepts. Indeed, suppose A and B are mutually exclusive, then $P(A \cap B) = 0$ which would not, in general, mean A and B were independent.

Conversely, supposing A and B are independent, then, in general, $P(A)P(B) \neq 0$, so that $P(A \cup B) \neq P(A) + P(B)$. Therefore, mutual exclusivity and independence are logically inequivalent notions.

More generally, we can define the independence of a sequence of events in an experiment. We shall desire such a definition below in chapter 6 for stochastic processes. As such, we provide it here.

Definition 26. *Let $A_1, A_2, \ldots, A_r \in \beta$ be a sequence of events in an experiment Ω. We say the sequence of events is independent if both*

$$P(\cap_{i=1}^r A_i) = \prod_{i=1}^r P(A_i)$$

and, for each $1 \leq i, j \leq r$ such that $i \neq j$, we have

$$P(A_i \cap A_j) = P(A_i)P(A_j)$$

Let us consider a few routine consequences of the definition in the following proposition.

Proposition 12. *Let $A, b \in \beta$ be independent events in an experiment Ω. Then the following events are also independent.*

1. *A and B^c*

2. *A^c and B*

3. *A^c and B^c*

To prove this proposition, let us begin by proving the first item. To that end, recall from chapter 1 that we can write $A = (A \cap B) \cup (A^c \cap B)$, so that

$$\begin{aligned}
P(A) =& P(A \cap B) + P(A^c \cap B) \\
P(A) - P(A \cap B) =& P(A \cap B^C) \\
P(A) - P(A)P(B) =& \\
P(A)(1 - P(B)) =& \\
P(A)P(B^c) =& P(A \cap B^c)
\end{aligned}$$

which illustrates the first item by the definition of independence. The proof of the second item is the same *mutatis mutandi*. The third claim is an application of DeMorgan's law. Observe

$$\begin{aligned}
P(A^c \cap B^c) =& P((A \cup B)^c) \\
=& 1 - P(A \cup B) \\
=& 1 - (P(A) + P(B) - P(A \cap B)) \\
=& 1 - (P(A) + P(B) - P(A)P(B)) \\
=& (1 - P(A))(1 - P(B)) \\
=& P(A^c)P(B^c)
\end{aligned}$$

which again exhibits the complements of independent events are independent themselves by the definition of independence. Let us now consider several examples that require us to compute the probability of independent events before considering the general case of "dependent events" or events whose probabilities depend conditionally upon one another, below.

First, consider the experiment of tossing an unfair coin twice, where $P(H) = \frac{6}{10}$. The simple events of this experiment are given by pairs $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$. Clearly the second flip is independent of the first. Let $A, B \in \beta$ be the events "Heads first" and "Heads second," respectively. What is the probability of the event, "two Heads"?

The solution requires us to compute the probability of the simple event (H,H). Observe, $P(H,H) = P(A \cap B)$, as they are described above. So, given these events are independent, we have that $P(H,H) = P(A)P(B) = \left(\frac{6}{10}\right)^2$.

Continuing, what is the probability of event "one Head"? To answer this question, let us character this even, say C, in terms of A and B, as above. As such, write $C = (A \cap B^c) \cup (A^c \cap B)$. Now the sets in this union are disjoint, so that $P(C) = P(A \cap B^c) + P(A^c \cap B)$. Now, as A and B are independent events, so too are the intersections of events in the arguments of P by the earlier proposition. Therefore, we have $P(C) = \frac{6}{10}\frac{4}{10} + \frac{4}{10}\frac{6}{10} = \frac{12}{25}$.

As we continue, let us compute the probability of the event "no Heads". Again, we shall describe this event in terms of A and B, as above. Hence, let $D = A^c \cap B^c$ and notice that, once more, A^c and B^c are independent by the proposition. Thefore, $P(D) = P(A^c)P(B^c) = \left(\frac{4}{10}\right)^2$. And last, let us compute the probability of the event "at least one Head". We can achieve this indirectly by recognizing that this event, say E, is indirectly given in terms of A and B by $E = D^c$, so that $P(E) = 1 - P(E^c) = 1 - P(D) = 1 - \frac{16}{100}$.

Next, let us consider the experiment involving tossing two fair four-sided dice. The simple events of this experiment consists of pairs $\Omega = \{(i,j) \mid 1 \leq i,j \leq 4\}$. Notice that, by the multiplication rule, $|\Omega| = 16$. Moreover, by the fairness hypothesis, the simple events are equally as likely when we adopt the canonical definition of P, so that $P(i,j) = \frac{1}{16}$. Equivalently, Ω is an equally-likely experiment.

Consider the events "the first roll is i" and "the second roll is j" denoted $A(i)$ and $B(j)$, respectively. Our first question is whether $A(i)$ and $B(j)$ are independent for admissible i,j.

3.1 WORKED EXAMPLES

1. 3.1 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

2. 3.1 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3. 3.1 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3.2 Conditional Probability and Baye's Theorem

Let us now consider the general case of the joint probability of a sequence of events in an experiment and see how an intrinsic characterization of independent events emerges from this consideration. The joint probability of two events is most clearly expressed by considering the probability of an event A given another, in general distinct, even B has occurred. Perhaps more abstractly, we have to consider the probability of A *relative* to B or what happens when Ω is replaced by B. Trivially, $P(A) = \dfrac{P(A \cap \Omega)}{P(\Omega)}$. More generally, however, replacing Ω by B immediately gives rise to the following definition.

Definition 27. *Let $A, B \in \beta$ be events in an experiment Ω. We say the conditional event of A with respect to B is the intersection of A and B as a subset of B, denoted $A|B$. Furthermore, the conditional probability of A given B has occurred is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Plainly this definition arises from replacing Ω by B in the computation of $P(A)$ which, pendantically, one could say now is the conditional probability of A given Ω has occurred. Moreover, hopefully this relative view of the probability of an event clarifies the meaning of independent events. Indeed, two events are independent if replacing the sample space by one of them does not alter the likelihood of the other. Indeed, supposing A and B are independent, then $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(A)P(B)}{P(B)} = P(A)$. Indeed, a well-known alternative to our definition of independence is that A and B are independent events if $P(A|B) = P(A)$ and *vice versa*. One must be careful to specify this alternative definition only applies to $B \neq \emptyset$.

This last point is often raised to justify defining conditional probability without implicitly assuming that B is not the empty set. In this case, one defines conditional probability through the so-called multiplication rule for probabilities, which is $P(A|B)P(B) = P(A \cap B)$. This reformulation of conditional probability is the basis for induction used to prove the following generalization, describing the joint probability of an arbitrary finite sequence of events.

Theorem 13. *Let $A_1, A_2, \ldots, A_k \in \beta$ be a sequence of events in an experiment Ω. Then*

$$P(\cap_{i=1}^{k} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3) \cdots P(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1})$$

Now write some examples

Last we state and prove a most famous theorem which computes the *a posteriori* probability of a cell in a partition of an experiment. The following result is known as Baye's Theorem.

Theorem 14. *Let (Ω, β, P) be a discrete probability space together with a partition of length r, say $A_1, A_2, \ldots, A_r$. Then for an arbitrary event $B \in \beta$, we have*

$$P(A_i|B) = \frac{P(A_i)(P(B|A_i))}{\sum_{j=1}^{r} P(A_j)P(B|A_j)}$$

for any $1 \leq i \leq r$.

The proof is fairly easy as it is an application of the definition of conditional probability and the law of total probability. Indeed, the numerator is given by $P(A_i \cap B) = P(A_i)P(B|A_i)$. The denominator is similar, for the law of total probability gives that $P(B) = \sum_{j=1}^{r} P(B \cap A_j)$ and alike the computation for the numerator, we have, for each j, that $P(B \cap A_j) = P(A_j)P(B|A_j)$, as desired.

3.2 WORKED EXAMPLES

1. 3.2 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

2. 3.2 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

   (c) $C =$

   We proceed by

   Solution (a) A.

   Solution (b) B.

   Solution (c) C.

3. 3.2 Determine the probability the following events:

   (a) $A =$

   (b) $B =$

(c) $C =$

We proceed by

Solution (a) A.

Solution (b) B.

Solution (c) C.

3.3 Chapter 3 Exercises

1. Prove that, for any probability space $(\Omega, \mathscr{B}, P)$ and events $A, B \in \mathscr{B}$, if

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

then A and B are independent.

rmk: the"proof" here is just a manipulation of the given equality to obtain the definition of independence, just like the"proofs" in class. Remember, too, that you are allowed to assume the centered identity, so to demonstrate/prove that the conclusion"A and B are independent" from it, you merely have to derive the definition of independence given the centered identity.

2. Consider the following experiment. A playing card is drawn from a deck of 52 cards and replaced, then a second card is drawn. The associated sample space Ω consists of pairs which record the suit and denomination of the two cards drawn. Let A be the event, "the first card is a spade. Let B be the event, "the second card is a spade." Let C be the event, "both cards have the same color." Please recall that playing cards are either red, the hearts and diamonds, or black, the spades and clubs.

Determine whether

a.) A and B are independent.
b.) B and C are independent.
c.) A, B, and C are independent.

rmk: Note that to solve c.) you will need to use the general definition of independence we gave in class, because there are three events whose independence of which you are trying to determine.

3. Consider the following experiment of rolling both one six sided red die and one six sided blue die and recording the individual outcomes. What is the probability of the event the sum of the numbers on both the red and the blue dice is 7 and the value on the blue die is larger than the value on the red die?

4. Suppose an urn contains 25 red balls and 15 blue balls. Consider the experiment of choosing 2 balls from the urn, without replacement. The associated sample space consists of the pairs recording the color of the balls selected.

   What is the probability of the event, "both balls are red?"

5. A test that screens for illegal drug use, i.e. a“drug test," is used in a large population of people in which 4 % of who actually use drugs. Suppose that the false positive rate in the drug test is 3 % and that the false negative rate in the drug test is 2 %. Accordingly, an individual who actually uses drugs tests positive for drug use 98 % of the time, whereas an individual who does not use drugs tests negative for drug use 97 % of the time. What is the probability that an individual randomly chosen who tests positive for illegal drug use actually uses illegal drugs?

6. A laboratory blood test is 95 % effective in detecting a certain disease when it is actually present. However, the test also yields a “false positive" result for 1 % of the healthy persons tested. If .5 % of the population actually has the disease, what is the probability a person has the disease given the test result is positive.

4 Discrete Random Variables

In this chapter we introduce discrete random variables in the generality of probability theory, which is to say, we introduce their general features and postpone examples until the next chapter. A random variable is a powerful computational tool inasmuch as it is a formalization of how one may characterize experiments by numerical parameters. Most conveniently, we can compute the probability of events in the experiment according to the numerical parameter assigned to it by the random variable. This technique is so effective it is often suppressed in introductory courses on this subject by teaching the topic as though it were self evident. Yet, in this chapter, we explore this topic at the level of rigor appropriate to an undergraduate course.

4.1 Partitions *qua* Equivalence Relations

This section is intended to help explain what one means by the probability of the value of a random variable, or plainly speaking, a number. We provide the correct insight into how such a notion is obtained, as opposed to the informal treatment in comparable undergraduate textbooks. Curiously enough, we begin with a discussion on how the notion of the equality elements in a set may be handled in the present generality. It is curious indeed that such a notion is not as insipid as a layperson might suggest it is. As usual, good mathematics proceeds from a good definition.

Definition 28. *Let S and T be sets, then we say a relation R between S and T is any subset of their Cartesian product. To wit, $R \subset S \times T$ is a relation. Furthermore, we say $s \in S$ is related by the relation R to $t \in T$ and vice versa if $(s,t) \in R$. We denote the ordered pairs of a relation R by sRt. In the special case when $S = T$ we simply say that R is a relation on S.*

Upon consideration this definition may appear to be a mere triviality, for it seems to only impart a name to what is otherwise an ordinary subset of a Cartesian product. However, the sense in which it specifies a relation between components of an ordered pair is by specifying a membership condition. Let us consider what is perhaps the most famous relation of all, the set of rational numbers, $\mathbb{Q}$. Indeed, by definition, this set is defined to be ratios of integers $\dfrac{m}{n}$ such that $n \neq 0$. Equivalently, $(m,n) \in R = \mathbb{Z} \times \mathbb{Z}^*$, where $\mathbb{Z}^* = \mathbb{Z} \setminus 0$.

One notices immediately that this description of rational numbers as the aforementioned Cartesian product seems lacking, for it neglects the notion of equivalent fractions. This observation motivates the next definition.

Definition 29. *We say a relation R on a set S is an equivalence relation if it satisfies the following three axioms. Let $s_1, s_2, s_3 \in S$, then R must satisfy*

1. *R is reflexive, viz. $s_1 R s_1$*

2. *R is symmetric, viz. $s_1 R s_2$ and $s_2 R s_1$*

3. *R is transitive, viz. if $s_1 R s_2$ and $s_2 R s_3$, then $s_1 R s_3$*

The existence of an equivalence relation on a set S imparts a new structure to its elements, namely, they can be organized into subsets defined as collections that are equivalent with respect to the relation R. Specifically, if R is an equivalence relation on S then we say the equivalence class of $s \in S$ with respect to R, or less formally, its equivalence class, is $\mathscr{E}_s = \{s' \in S \mid sRs'\} \subset S$. That is, the subset of elements in S equivalent to s with respect to R. Now the structure these equivalence classes impart to the elements of S is a familiar one, for it is that of a partition.

Theorem 15. *Let R be an equivalence relation on S, then the set of equivalence classes with respect to R*

$$\{\mathscr{E}_s\}_{s \in S}$$

induce a partition on S. Conversely, given a partition of S, say $A_1, A_2, \ldots, A_r$, then the relation R on S defined by

$$sRs' \text{ if and only if } s, s' \in A_i$$

for $1 \leq i \leq r$ is an equivalence relation.

First, we note, however obvious, that $s \in \mathscr{E}_s$, so that every element of S is contained in some equivalence class. Therefore, $S = \cup_{s \in S} \mathscr{E}_s$. Next, so show that the set of equivalence classes is partition, we must show that the classes are pairwise-disjoint. So this end, we shall suppose that $a \in \mathscr{E}_s \cap \mathscr{E}_{s'}$ for some $a \in S$. Then, by definition, both aRs and aRs', so that sRs', as R is transitive. Thus $\mathscr{E}_s = \mathscr{E}_{s'}$ or they are disjoint. Accordingly, the set of equivalence classes $\{\mathscr{E}_s\}_{s \in S}$ is a partition of S. The proof of the converse to this statement in the theorem is routine.

Now we shall end this section with the result germane to our motivation, that is, to rigorously define below in the next section the probability of a value of a random variable. So let $f : S \to T$ be a function of sets. Observe that the function f defines a equivalence class on its domain, S. Indeed, we shall define two elements in S to be equivalent if their images under f are the same. That is, define an equivalence relation R on S by

$$sRs' \text{ if and only if } f(s) = f(s')$$

Clearly R is both reflexive and symmetric. That R is also transitive is just as obvious, but we spell this out anyway by writing when both s_1Rs_2 and s_2Rs_3 or $f(s_1) = f(s_2)$ and $f(s_2) = f(s_3)$ then s_1Rs_3 or $f(s_1) = f(s_3)$. We say this equivalence relation is the relation on S induced by f. This motivates another important definition through out mathematics, and again, one that we shall use below.

Definition 30. *Let $f : S \to T$ be a set function and R the relation on S induced by f. Then we say the equivalences classes with respect to this relation are the fibres of f. Furthermore, we denote the fibres of f by*

$$\mathscr{E}_s = \{f = t\}$$

where $\{f = t\} = \{s \in S \mid f(s) = t\}$

Notice that the fibres of f are the same the pre-images of elements in T contained in the image of f. The manner in which a function fibres its domain is an important insight intellectually and speaks to the ontological role that functions play in mathematics. Indeed, many great mathematicians of the previous century made great strides in the subject through their appreciation of this principle. Yet, beside whatever great strides we are alluding to may be, there is an elementary result, now at hand, we shall use below to compute the probability of values of a random variable.

Theorem 16. *Let $f : S \to T$ be a function of sets. Denote by S_f the set of fibres of f. Then there is a bijection, say ϕ, from S_f to $f(S)$.*

To prove this theorem, let us define a map $\phi : S_f \to f(S)$ by $\phi(\{f = t\}) = t$. Suppose now for $t, t' \in f(S)$ that $t \neq t'$, then we have $\{f = t\} \cap \{f = t'\} = \emptyset$. Hence ϕ is well-defined. To show ϕ is injective, note that $t = t'$ implies $\{f = t\} = \{f = t'\}$. Surjectivity is just as straightforward, for f necessarily surjects onto its image. In other words, by definition, for $t \in f(S)$ there exists an $s \in S$ such that $f(s) = t$. As the fibres of f partition S, there is a fibre containing s, say $\{f = t\}$. Then, by construction, $\phi(\{f = t\}) = t$, showing *phi* is a surjection and therfore, a bijection.

An application of this theorem explains how to incorporate the elementary concept of equivalent fractions into our initial description of the rational numbers as a relation R on $\mathbb{Z} \times \mathbb{Z}^*$. Indeed, define $f : R \to \mathbb{Q}$ $f(m, n) = \dfrac{m}{n}$. Then $f(R) = \mathbb{Q}$ and by the theorem, there is a bijection between the fibres of f and $\mathbb{Q}$. The elements of a fibre of f are equivalent fractions. As such, one can now appreciate how the fractions of elementary school are a first exposure to equivalence classes.

This application concludes the analogy between set theoretic operations and arithmetic ones. Whereas unions correspond to sums, intersections to products, and complements to difference, one could say that equivalence relations correspond to division. It is best to think of this operation as dividing a set or forming quotients of its elements by regarding equivalence classes as generalizations of fractions. Indeed, this way of thinking continues on in other subjects of mathematics, such as forming quotient groups in group theory, or of course, quotient stacks in algebraic geometry.

4.2 Discrete Random Variables

Broadly speaking, a random variable is a numerical summary of an experiment. We express such a summary of an experiment formally in terms of a discrete probability space together with a real-valued function, that is to say an assignment of simple events characterized by numbers of the summary to the very numbers by which they are characterized. This is the concept of a discrete random variable, if one remembers to think of assignments as functions.

Definition 31. *Let (Ω, β, P) be a discrete probability space. Then we say a real-valued function*

$$X : \Omega \to \mathbb{R}$$

such that $X^{-1}(x) \in \beta$ for $x \in X(\Omega)$ is a discrete random variable.

A few comments on this fundamental definition are in order. First, notice that the image of X is itself discrete. Indeed, by definition, any function surjects onto its image, and by chapter 1, this cardinality of Ω is an upper bound for the cardinality of its image, denoted Ω_X by either an abuse of notation or with the theorem of the previous section in mind. As such, Ω_X is a discrete set. Second, and perhaps most importantly, the condition that $X^{-1}(x) \in \beta$ means that for all $x \in \Omega_X$, we have the set of pre-images is a subset of the σ-algebra for which X is defined. Recall the notion for this set from the previous section; *mutatis mutandi*, we write Ω_X for the set of fibres of X. Accordingly, the aforementioned condition can be read to mean that $\Omega_X \subset \beta$. Third, as we shall only consider discrete random variables in this text, we shall often ignore the adjective "discrete" in our exposition. Moreover, one assumes there exists a discrete probability space for which X is defined if one is not mentioned explicitly.

Now the summary of an experiment furnished by a random variable X is insufficient for our purposes of simplifying the probability of events in an experiment in terms of such a summary alone. Indeed, X merely assigns real-numbers to events in an experiment. In order to elevate the addition of such a numerical summary X to something that can express the probability of an event it summarizes in terms of the probability theory we have studied, we shall introduce a new probability space, that of the *induced probability space*, defined below.

So, consider an arbitrary discrete random variable X with respect of a discrete probability space (Ω, β, P). We want to produce from this data a new triple (Ω_X, β_X, P_X), which we will call the induced probability space. In order to do this, as the notation suggests, we may take the fibres of X, or, equivalently, its image, as the sample space for this triple. As we observed above, Ω_X is a countable set, by construction, therefore, by chapter 1, we adopt as the induced σ-algebra the canonical choice. Namely, $\beta_X = 2^{\Omega_X}$, the power set of Ω_X. Lastly, the reason for our digression in the first section of this chapter into equivalence relations, is to defined P_X in such a way that it satisfies Kolmogorov's axioms.

Indeed, by the definition of X, we have the set of its fibres is a subset of the σ-algebra, β. Accordingly, we define, for $\{X = x\} \in \beta$ the probability function P_X by the formula

$$P(\{X = x\}) = P_X(x)$$

We must emphasize this definition is only possible as the set of fibres of X is a subset of β, the domain of the probability function P. It is in this manner that a random variable induces a probability, provided that P_X satisfies Kolmogorov's axioms. That is the content of the next theorem.

Theorem 17. *Let (Ω, β, P) be a discrete probability space and $X : \Omega \to \mathbb{R}$ a random variable. Then the function $P_X(x) = P(\{X = x\})$ satisfies Kolmogorov's axioms.*

To prove this theorem, we proceed in the routine manner by verifying the three axioms. First, $P_X(\Omega_X) = \sum_{x\in\Omega_X} P_X(x) = \sum_{\{X=x\}} P(\{X = x\})$. Now, as the fibres of X partition Ω, we can write

$$\Omega = \bigcup_{u\in\Omega}\{X = x\}$$

As P satisfies Kolmogorov's axioms, we have $1 = P(\Omega) = P(\bigcup_{u\in\Omega}\{X = x\}) = \sum_{\{X=x\}} P(\{X = x\})$. Therefore, P_X satisfies the first axiom.

Next, to prove that P_X satisfies the second axiom, we merely notice that $P_X(x) = P(\{X = x\}) \geq 0$ as P itself satisfies Kolmogorov's second axiom.

Last, to prove that P_X satisfies the third axiom, we have $P_X(\bigcup_{i\in I} x_i)$, where I is some finite set of indices. Then, by definition, $P_X(\bigcup_{i\in I} x_i) = P(\bigcup_{u\in X^{-1}\left(\bigcup_i x_i\right)}\{X = x_i\}) = \sum_{u\in X^{-1}\left(\bigcup_i x_i\right)} P(\{X = x_i\}) = \sum_{\{X=x_i\}} P_X(x_i)$ where the middle equality follows from the fact that P itself satisfies Kolmogorov's third axiom. It should be obvious that a discrete set of real-numbers comprise a pairwise-disjoint sequences of subsets of $\mathbb{R}$. Altogether, our discussion here proves the following theorem and definition.

Theorem 18. *Let (Ω, β, P) be a discrete probability space and $X : \Omega \to \mathbb{R}$ a random variable. Then the triple (Ω_X, β_X, P_X) defined above is a discrete probability space. We say this probability space is the induced probability space. Furthermore, to unburden notation, hereafter, we denote the induced probability function $P_X(x) = p(x)$, unless otherwise noted.*

Introducing random variables to the study of the likelihood of events in an experiment does more than furnish a convenient numerical summary which can be used to compute probability. They also furnish such experiments with additional insightful details, and the study of these is what follows.

4.3 The Histogram and Examples

In this section, we shall introduce the first additional detail a random variable imparts to an experiment and also present some examples. So, let X be a random variable defined with respect to some discrete probability space. As noted in the previous section, Ω_X is a discrete subset of $\mathbb{R}$. Let us suppose further for the sake of simplicity that $|\Omega| = N$, for this supposition affords us index labels for the values of X as it is discrete. So we write $\Omega_X = \{x_1, x_2, \ldots, x_N\}$ and it is straightforward to see that Ω_X is in bijective correspondence with the set $[n] = \{1 < 2 < \cdots < n\}$ by $x_i \mapsto i$ for $1 \leq i \leq n$. We have enough notation for the following definition.

Definition 32. *Let X be a random variable defined with respect to a discrete probability space. Define h_X the real-valued piece-wise continuous function by the formula $h_X(x) = p(x_i)$ for $x \in [i, i+1)$ and $i \in [n]$. Then the graph of h_X in $\mathbb{R}^2$ is called the histogram of X.*

One notices that the area of the planar region determined by Γ_{h_X}, that is, the histogram of X, is equal to 1 by Kolmogorov's first axiom. Beside this interesting feature, the histogram allows one to visualize an experiment as a system of particles $[n]$ together with weights $p(x_i)$. We shall pursue this observation more substantially below when we discuss the expectation of a random variable as well as its variance. For now, however, we present a littany of examples.

4.4 Generating Functions, Transformations, and the Law of the Unconscious Statistician

In this section we introduce formal power series and take the generating function associated to a random variable X as our main example. We use this object in conjunction with the concept of the transformation of a random variable, which shall also be introduced herein, to prove the theorem of the unconscious statistician. The theorem is useful throughout the subject, but in general, and in particular for the families of discrete random variables we present in chapter 5.

So, a *formal power series* is an element of the set $\mathbb{R}[[z]] = \{\sum_x^\infty r_x z^x\}$, where $r_x \in \mathbb{R}$ for all x. What distinguishes these power series from those of say, an elementary course in Calculus, is that the notion is convergence is ignored, whence the adjective "formal." Nonetheless, this set is endowed with the same structure it has when convergence is maintained, that is to say, one may add and subtract or multiply and divide elements as usual. Furthermore, one also has the operation of formal differentiation which also ignores the convergence of limits. To wit, one defines an operator $\dfrac{d}{dz}$ on this set in the obvious way. Let $f(z) = \sum_x r_x z^x \in \mathbb{R}[[z]]$, then

$$\frac{df(z)}{dz} = \textstyle\sum_x r_x x z^{x-1}$$

defines the formal derivative operator on the algebra $\mathbb{R}[[z]]$. We shall use this formula below to define the both the formal expectation and the formal variance of a random variable. It must be emphasized that, although this definitions are free of any context, the formulas they produce are most useful and are the ones we still use in applications.

We shall be interested in these objects when $r_x = p(x)$ for some random variable X. Indeed, as polynomials are distinguished by their coefficients, so too are formal power series. In particular, when they are probabilities of the values of a random variable, we arrive at the notion of a generating function. The following definition provides all of the details.

Definition 33. *Let X be a random variable with respect to some discrete probability space. Then we say the formal power series $g(z)$ such that*

$$g(z) = \textstyle\sum_{x \in \Omega_X} p(x) z^x$$

is the generating function of X.

Again it must be emphasized that the function generates X in the sense that the sequence of its coefficients completely determine X. We shall take advantage of this feature in conjunction with the formal derivative to obtain to important characteristics of the random variable X.

Definition 34. *Let X be a random variable and $g(z)$ its generating function. Then we say*

1. $E(X) = \frac{dg(z)}{dz}|_{z=1} = g'(1)$ *is the expectation of X.*
2. $V(X) = \frac{d^2g(z)}{dz^2}|_{z=1} + \frac{dg(z)}{dz}|_{z=1} - \left(\frac{dg(z)}{dz}|_{z=1}\right)^2 = g''(1) + g'(1) - (g'(1))^2$ *is the variance of X.*

Although we have yet to define the features of an experiment, we shall recognize their definitions later in the following computations. First, given the generating function $g(z)$ of a random variable X, we have, computing its formal derivative

$$g'(z) = \textstyle\sum_{x \in \Omega_X} x p(x) z^{x-1}$$

so that, evaluating at $z = 1$, we have

$$g'(1) = \sum_{x \in \Omega_X} x p(x)$$

we obtain the definition of the expectation of a random variable X given below.

A similar computation shows that $E(X^2) = g''(1) + g'(1)$, so that, $V(X) = E(X^2) - (E(X))^2$. This second computation is mysterious for two reasons. First, it is not perfectly clear what $E(X^2)$ means. However, it is not hard to imagine that if we write $Y = X^2$, then it simply means the first derivative of the generating function of the random variable Y. Yet, how this is expressed in terms of the original or independent random variable X shall motivate the following discussion on *transformations* of random variables. The second aspect that shall remain mysterious is why $V(X)$ can be simplified in terms of a linear combination of expectations. We shall pursue this point in subsequent sections and for now content ourselves with this most useful formula.

The the definition of variance suggests, we must make sense of $E(Y)$ for a random variable $Y = f(X)$ that depends on a random variable X. Such a dependent random variable is referred to as a transformation of the random variable X according to curious habit of statisticians to rename mathematical conventions in their own terms. However, conceptually, it should be emphasized that a transformation of a random variable is merely a random variable that is a function of another according to the standard notions of elementary mathematics.

Given a transformation Y of a random variable X, it too is a random variable, most obviously, with respect to the induced probability space of X. Namely, $f : \Omega_X \to \mathbb{R}$, where $f(x) = y$ for $x \in \Omega_X$, defines with obvious notation, a random variable Y with respect to the induced probability space of X. As above, the image of a discrete set is discrete, so that we may take $\Omega_Y = f(\Omega_X)$ as the induced sample space with respect to Y. Further, as Ω_Y is discrete, we also take the canonical σ-algebra to be $\beta_Y = 2^{\Omega_Y}$. So, as usual, the challenge is defining the probability function, P_Y.

We proceed to define P_Y *mutatis mutandi*. Indeed, the fibres of Y are the subsets $\{f = y\}$ or, equivalently, $\{x = f^{-1}(y)\}$ where f^{-1} denotes the right inverse of f. Such a right inverse exists as any function surjects onto its image. In any case, we can therefore define P_Y by the formula

$$\begin{aligned} P_Y(y) =& P_X(\{f = y\}) \\ =& \sum_{x \in f^{-1}(y)} P_X(x) \end{aligned}$$

We leave the proof this definition satisfies Kolmogorov's axioms to the reader, as it amounts to re-writing the proof that P_X does. We have therefore proven, with notation as above, that (Ω_Y, β_Y, P_Y) is a discrete probability space. We refer to this space, if there is occasion to, as an induced probability space as well, although one might suggest that the "induced induced" probability space is more consistent. The reason for this discussion is to prove the following well-known theorem, that of the unconscious statistician.

Theorem 19. *Let X be a random variable together with a transformation, say $Y = f(X)$. Then*

$$E(Y) = \sum_{x \in \Omega_X} f(x)p(x)$$

To prove this, we proceed by the definition. So, consider the generating function of Y and write

$$g(z) = \sum_{y \in \Omega_Y} p(y)z^y$$

Then we can deduce the formula of the theorem in the following steps:

$$\begin{aligned} g(z) &= \sum_{y \in \Omega_Y} p(f(x))z^{f(x)} \\ g'(z) &= \sum_{y \in \Omega_Y} f(x)p(f(x))z^{f(x)-1} \\ E(Y) &= \sum_{x \in \Omega_X} f(x)p(x) \end{aligned}$$

where the last step follows from the definition of $p(y) = p(f(x))$ described above. Apparently the name for this theorem is due to impression a statistician could use it while unconscious. We shall use it below to obtain several key facts in such an unconscious manner.

4.5 The Expected Value and Variance of a Discrete Random Variable

In the previous section we formally defined both the expectation and variance of a random variable via its generating function. This section is intended to convey how these aspects of a random variable express intuitive features of an experiment. To that end, let us begin to interpret the expectation of a random variable in terms of experiment to which it is associated via the discrete probability space it is defined with respect to.

Definition 35. *Let X be a random variable and (Ω_X, β_X, P_X) the induced probability space. Then we say*

$$E(X) = \sum_{x \in \Omega_X} xp(x)$$

is its expected value.

Of course this definition is that of the previous section, however, now it is intended to be understand in terms of (Ω, β, P) or indirectly, in terms of (Ω_X, β_X, P_X). Let us consider the special case $\Omega_X = \{x_1, x_2\}$ or in order to recognize in what sense we mean to understand the expected value intuitively as the average value. The average value of a random variable X should be the real number that balances the values of X in the sense that $x_1p(x_1) = x_2p(x_2)$. If one were to regard Ω_X as a physical system, then it would be a lever with end points x_1 and x_2 and $E(X)$ would be the coordinates along the lever of the fulcrum that balances these end points when they are weighted by $p(x_i)$.

Let us imagine that our lever lies along the $x-axis$ and let us write $E(X)$ as its fulcrum point. Then, we have, as a consequence of the fact $x_1p(x_1) = x_2p(x_2)$ that $(E(X)-x_1)p(x_1) = (E(X) - x_2)p(x_2)$ but then, solving for $E(X)$ gives

$$\begin{aligned}
(E(X) - x_1)p(x_1) =& (x_2 - E(X))p(x_2) \\
E(X)p(x_1) + E(X)p(x_2) =& x_1p(x_1) + x_2p(x_2) \\
E(X) =& \frac{x_1p(x_1) + x_2p(x_2)}{p(x_1) + p(x_2)}
\end{aligned}$$

where the denominator sums to 1 by Kolmogorov's first axiom. The upshot of this is that the fulcrum or balancing point on our "lever" Ω_X is $E(X) = x_1p(x_1)+x_2p(x_2)$. We may and shall therefore generalize this illustration to an "arbitrary lever" Ω_X, as above in our definition, and interpret the expected value as the fulcrum point. A more sophisticated perspective is to regard $E(X)$ as the *center of gravity* for this lever, but an adequate explanation of this perspective is outside of the scope of this text.

Next we intend to prove a famous linearity result, namely:

Theorem 20. *Let both X and Y be random variables defined with respect to the discrete probability space (Ω, β, P), then*

$$E(X \pm Y) = E(X) \pm E(Y)$$

However, to prove this result, we shall introduce multiple random variables or, perhaps one could say, random variables of several variables, however abusively. The up shot is they are random variables defined with respect to a probability space that depend on several random variables. In particular, we shall prove the above theorem for the random variable $Z = X + Y$ in conjunction with the aforementioned definition of expected value. We shall treat the case of two random variables carefully enough and state results that depend on n-random variables without proof. First we will briefly review double summation.

Let us consider the *double sum*, defined as follows. Consider some set theoretic function $z_{ij} = f(i, j)$ defined on a discrete rectangle $R = \{1 \leq i \leq m, 1 \leq j \leq n\}$. Then the double sum over the discrete rectangle R is the sum of the mn elements of the array z_{ij} determined by f. To wit,

$$\sum \sum_{(i,j) \in R} z_{ij}$$

Now, there is a analogue of Fubini's theorem for double sums which must be observed. It says that the double sum may be computed as an iterated sum. Assuming R is obtained as a Cartesian product of discrete intervals, say $R = I \times J$, then

$$\begin{aligned} \sum \sum_{(i,j) \in R} z_{ij} =& \sum_{i \in I} \left(\sum_{j \in J} f(i, j) \right) \\ =& \sum_{j \in J} \left(\sum_{i \in I} f(i, j) \right) \end{aligned}$$

where the right hand side is interpreted as an iterated sum, meaning to first sum over J to obtain a single dependent upon i and then to compute the sum over I second and *vice versa*. We will use this formula below to simplify the sum over the image of a point-wise combination of random variables into iterated sums. Beyond this application, it is useful to prove several properties of expected value and variance.

Pursuant to the first application, let X and Y be random variables defined with respect to the discrete probability space (Ω, β, P). Observe that both X and Y furnish maps to $\mathbb{R}$, by definition, so together they are the *coordinate functions* of the map $X \times Y : \Omega \to \Omega_X \times \Omega_Y \subset \mathbb{R}^2$, where $X \times Y(u) = (X(u), Y(u))$ for $u \in \Omega$. Given this data, we obtain a new random variable $Z : \Omega \to \mathbb{R}$ by $Z(u) = f(X(u), Y(u))$, where $f : \mathbb{R}^2 \to \mathbb{R}$. The new random variable Z satisfies the defining condition by observing that $\{Z = z\} = \{X = x\} \cap \{Y = y\}$. Accordingly, we obtain the induced probability space with respect to Z, the *multiple random variable*, where Ω_Z is the image of the discrete space $\Omega_X \times \Omega_Y$ under f and the probability function $P_Z(z) = p_{XY}(x, y)$.

The last item is significant for several concepts below, and we shall refer to it as the *joint probability function*, denoted $p_{XY}(x, y)$ no matter what the specific formula for f is.

Moreover, for emphasis, we shall refer to (Ω_Z, β_Z, P_Z) as the *joint induced probability space.* Associated to this data are the *marginal probability functions.* Indeed, we are able to recover the individual probability functions of X and Y from the joint probability function.

Definition 36. *Let $Z = f(X, Y)$ be a multiple random variable with respect to the discrete probability space (Ω, β, P). Then we say*

1. $p_X(x) = \displaystyle\sum_{y|p(x,y)>0} p_{XY}(x, y)$ *is the marginal probability function with respect to X*

2. $p_Y(y) = \displaystyle\sum_{x|p(x,y)>0} p_{XY}(x, y)$ *is the marginal probability function with respect to Y.*

Observe that the marginal probability functions recapitulate the corresponding probability functions. Indeed, with notation as above, $\sum_{x\in\Omega_X}\left(\sum_{y|p(x,y)>0} p_{XY}(x, y)\right) = P(\Omega_X) = 1$, for example. Now, to prove this theorem, observe that we can write the definition of the expectation of Z in terms of the marginal probability functions by defining the multiple random variable $Z = f(X, Y) = X + Y$. Then, we have

$$\begin{aligned}
E(Z) = &\sum_{z\in\Omega_Z} zp(z) \\
= &\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} (x + y)p_{XY}(x, y) \\
= &\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} xp_{XY}(x, y) + \sum_{y\in\Omega_Y}\sum_{x\in\Omega_Y} yp_{XY}(x, y) \\
= &\sum_{x\in\Omega_X} xp_X(x) + \sum_{y\in\Omega_Y} yp_Y(y) \\
= &E(X) + E(Y)
\end{aligned}$$

as desired. An immediate corollary is the following theorem.

Theorem 21. *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables defined with respect to the discrete probability space (Ω, β, P) and $Z = \sum_{i=1}^n X_i$ be the corresponding multiple random variable, then*

$$E(Z) = \textstyle\sum_{i=1}^n E(X_i)$$

We continue with a few other easy properties, with the goal to collectively demonstrate that E is a linear operator on the vector space of random variables defined with respect to a discrete probability space.

Theorem 22. *Let $c \in \mathbb{R}$ be a constant, then $E(c) = c$*

The proof of this theorem is immediately from the definition of expectation. Indeed, treating the constant c as a constant random variable, we have $E(c) = \sum_{x\in\Omega_c} xp(x)$, but $\Omega_c = \{c\}$ so that $E(c) = c$, since $p(c) = 1$, as desired. Next, to complete demonstrating the linearity of E as an operator, we have the following theorem.

Theorem 23. *Let X be a random variable and $c \in \mathbb{R}$ a constant, then $E(cX) = cE(X)$.*

Again, we may appeal to multiple random variables to prove this theorem. Define $Z = f(X, c) = cX$, then we have, by definition

$$\begin{aligned}E(Z) =& \sum_{z \in \Omega_Z} zp(z) \\ =& \sum_{x \in \Omega_X} \sum_{c} cxp_{cX}(x, c) \\ =& \sum_{x \in \Omega_X} \sum_{c} cxp_X(x) \\ =& \sum_{x \in \Omega_X} cxp_X(x) \\ =&c \sum_{x \in \Omega_X} xp(x) \\ =&cE(X)\end{aligned}$$

as desired. So, we have the following corollary to the above results.

Theorem 24. *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables defined with respect to the discrete probability space (Ω, β, P), $c_1, c_2, \ldots, c_n \in \mathbb{R}$ a sequence of random variables, and $Z = \sum_{i=1}^{n} c_i X_i$ be the corresponding multiple random variable, then*

$$E(Z) = \textstyle\sum_{i=1}^{n} c_i E(X_i)$$

that is, expected value is a linear operator.

Now we may introduce the next significant aspect of a random variable invoked to understand experiments quantitatively, that of its variance, which follows from our interpretation of $E(X)$ as the value of X that balances its values. Indeed, according to this interpretation, we shall measure how spread out the values of X are about $E(X)$: this measurement is the variance of X.

So, to derive the variance from the point of view that it measures how spread out the values of X are about $E(X)$, consider more generally the distance between X and an arbitrary constant, say c. In particular, write $d(X, c) = (X - c)^2$ for this distance. Now, let us compute the value c that minimizes the expected value of this distance function, d, which we think of as the transformation of the random variable, X. Then, by the law of the unconscious statistician, we have $E(d(X, c)) = \sum_{x \in \Omega_X} (x - c)^2 p(x)$. Moreover, by properties of expectation, we have that

$$\begin{aligned}E(d(X, c)) =&E(X - E(X) + E(X) - c)^2 \\ =&E(X - E(X))^2 + (E(X) - c)^2 + 2E((X - E(X))(E(X) - c))\end{aligned}$$

In general $E(X - E(X)) = E(X) - E(E(X)) = E(X) - E(X) = 0$ by properties of expectation, so the term on the right of the last equation vanishes. Therefore, if we wish to minimize $E(d(X, c))$ we must minimize only $E(X - E(X))^2 + (E(X) - c)^2$. The value $E(X - E(X))$ is arbitrary, however, we can still minimize $d(X, c)$ when $c = E(X)$, for $(E(X) - c)^2$ will then vanish. Accordingly, we have the following definition.

Definition 37. *Let X be a random variable and (Ω_X, β_X, P_X) the induced probability space. Then we say*

$$V(X) = \sum_{x \in \Omega_X} (x - E(X))^2 p(x)$$

is its variance. Furthermore more, with the same notation, we define the square root of the variance to be the standard deviation of X, denoted σ_X

This aspect of X is named appropriately, for a large value of $V(X)$ would correspond to the average distance of a value of X from its center of mass to be large whereas a small value would correspond to this average distance being small. At either extreme, the plain language description comports with our intuitive notion of variance in an experiment. We shall now enumerate the properties of $V(X)$.

Theorem 25. *Let X be a random variable and $V(X)$ its variance, then*

$$V(X) = E(X^2) - (E(X))^2$$

We shall refer to the proof of this theorem in the previous section on generating functions. It is most convenient in applications and proofs to remember this formula. Next, unlike the expected value of a random variable, the variance is most certainly not a linear operator.

Theorem 26. *Let X be a random variable and $c \in \mathbb{R}$ a constant. Then*

1. $V(c) = 0$
2. $V(cX) = c^2 V(X)$

Proving both items follows from the previous theorem with respect to the constant random variable, say $X = c$. Then we have $V(c) = E(c^2) - (E(c))^2 = c^2 - c^2 = 0$. Next, define $Y = cX$, then by definition, $V(Y) = E(c^2X^2) - (E(c))^2 = c^2E(X^2) - c^2 (E(X))^2 = c^2V(X)$, as desired. In the next section we will discover what hypotheses are required so that V behaves as an affine operator on a subspace of the vector space of linear operators defined with respect to a discrete probability space.

4.6 Independent Random Variables and the Coefficient of Correlation

In this section we explore further properties of the expected value and the variance of a random variable. In particular, under what condition is E multiplicative and V affine. This condition is an important property of sequences of random variables in its own right, and deserves to be singled out by a definition. Furthermore, it is one of the conditions we need to make sense of the probabilities of simple events in time dependent experiments.

Definition 38. *Let X and Y be discrete random variables, $Z = f(X,Y)$, and $p_{XY}(x,y)$ their joint probability distribution. Then we say X and Y are independent if*

$$p_{XY}(x,y) = p_X(x)p_Y(y)$$

for all $(x, y) \in \Omega_Z$

Under the auspices of this definition, we have the following proposition.

Proposition 13. *Let* X *and* Y *be independent random variables. Then the following statements are true:*

1. $E(X \cdot Y) = E(X) \cdot E(Y)$ *or* E *is multiplicative.*
2. $V(X + Y) = V(X) + V(Y)$ *or* V *is affine.*

We proceed by definition. Observe, for $Z = X \cdot Y$, we have

$$\begin{aligned} E(X \cdot Y) = & \\ =& E(Z) \\ =& \sum_{z \in \Omega_Z} z p(z) \\ =& \sum \sum_{(x,y) \in \Omega_{XY}} (xy) p_{XY}(x, y) \\ =& \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} x p_X(x) y p_Y(y) \\ =& E(X) \cdot E(Y) \end{aligned}$$

as desired. To demonstrate the second item, we rely upon the truth of the first. Observe,

$$\begin{aligned} V(X + Y) =& E(X + Y)^2 - (E(X + Y))^2 \\ =& E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\ =& E(X)^2 - (E(X))^2 + E(Y)^2 - (E(Y))^2 + 2E(XY) - 2E(X)E(Y) \\ =& V(X) + V(Y) + 2E(X)E(Y) - 2E(X)E(Y) \\ =& V(X) + V(Y) \end{aligned}$$

and therefore, V is affine.

Next, we generalize the independence of random variables by measuring their correlation. When discrete random variables are uncorrelated, we say that they are independent in the above sense. Consider the following measurement of how strongly two random variables are related to one another.

Definition 39. *Let* X *and* Y *be discrete random variables. Then we define their covariance to be*

$$CoV(X, Y) = E(X - E(X))E(Y - E(Y))$$

An interpretation of the covariance is that $X - E(X)$ and $Y - E(Y)$ measure the deviation of the variables from their expected value so that covariance is the expected value of the product of their deviations. GIVE INTERPRETATION

Proposition 14. *Let* X *and* Y *be discrete random variables. Then*

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

The proof is a practice in the definition of joint distribution.

$$\begin{aligned}
Cov(X,Y) &= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}(x-E(X))(y-E(Y))p_{XY}(x,y)\\
&= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}(xy - xE(Y) - yE(X) - E(X)E(Y))p_{XY}(x,y)\\
&=\sum_x\sum_y(xy)p(x,y) - \sum_x\sum_y xE(Y)p(x,y) - \sum_x\sum_y yE(X)p(x,y)\\
&\quad -\sum_x\sum_y E(X)E(Y))p(x,y)\\
&=E(XY) - E(Y)\sum_x xp_X(x) - E(X)\sum_y yp_Y(y) + E(X)E(Y)\\
&=E(XY) - (E(X)E(Y))^2 + E(X)E(Y)\\
&=E(XY) - E(X)E(Y)
\end{aligned}$$

as desired. This proposition allows us to define the coefficient of correlation.

Definition 40. *Let X and Y be discrete random variables. Then we say*

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

is the coefficient of correlation.

Notice that if X and Y are independent, then by the proposition $\rho_{XY} = 0$. However, if $\rho = 0$, then that does not necessarily mean that X and Y are independent. Accordingly, the converse of such a claim that uncorrelated variables are the same as independent ones is false. Other relationships between X and Y are now appreciable that covariance has been introduced.

Proposition 15. *Let X and Y be discrete random variables and $a,b,c,d \in \mathbb{R}$. Then the following statements are true:*

1. $\rho_{(aX+b)(cY+d)} = \rho_{XY}$
2. $-1 \leq \rho_{XY} \leq 1$
3. $\rho_{XY} = \pm 1$, *then* $Y = aX + b$

We begin by recognizing that $\sigma_{aX+b} = a\sigma_X$ since $V(aX+b) = a^2V(X)$, etc. Therefore, let us proceed by the proposition that simplifies the covariance formula

$$\begin{aligned}
\rho_{(aX+b)(cY+d)} &=\frac{E((aX+b)(cY+d)) - E(aX+b)E(cY+d)}{a\sigma_X c\sigma_Y}\\
&=\frac{acE(XY) - ac(E(X)E(Y))}{ac\sigma_X\sigma_Y}\\
&=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\\
&=\rho_{XY}
\end{aligned}$$

Next, let us define $\eta(t) = V(X)t^2 + 2Cov(X,Y)t + V(Y)$. Notice that $\eta(t) \geq 0$ since by unwinding the definitions of the coefficients shows that it can be written as the expected value of a non-negative discrete random variable. As such, it has at most one real root. Accordingly, its discriminant $\Delta = 4Cov(X,Y)^2 - 4V(X)V(Y) \leq 0$. This is equivalent to the inequality $-\sigma_X\sigma_Y \leq Cov(X,Y) \leq \sigma_X\sigma_Y$ or, by definition, $-1 \leq Cov(X,Y) \leq 1$ as desired.

Last, suppose that $\rho_{XY} = 1$, then that implies $\Delta = 0$. This in turn implies $\eta(t)$ has one real root t_0 of multiplicity two. This is true if and only if $P(\{(X - E(X))t_0 + (Y - E(Y))^2 = 0\}) = 1$ which in turn means $P(\{Y = aX + b = 1\}) = 1$ or $Y = aX + b$ with $a = -t_0$ and $b = E(X)t_0 + E(Y)$, ending the proof, for the case when $\rho_{XY} = -1$ is the same *mutatis mutandi.*

4.7 Chebyshev's Inequality

In this section we consider how the variance of a random variable determines the likelihood that the values of a random variable are far from its expected value. Indeed, according to our interpretation of the expected value of a random variable the center of gravity of a system of particles the intuition that a typical particle should be far away from the system's center of gravity is unlikely is made precise by Chebyshev's inequality.

Let us consider a discrete random variable X such $V(X)$ exists and $t \in \mathbb{R}_{\geq 0}$. Consider the probability of the equivalence class $P(\{|X - E(X)| \geq t\})$. The likelihood of this event measures how likely it is that X is at least t away from its expected value. By hypothesis, $V(X)$ exists and by definition we have $V(X) = \sum(X - E(X))^2$. Accordingly, we have

$$\begin{aligned} V(X) &= \sum_{u\in\Omega||X-E(X)|\geq t} (E - E(X))^2p(u) + \sum_{u\in\Omega||X-E(X)|<t} (E - E(X))^2p(u) \\ V(X) &\geq \sum_{u\in\Omega||X-E(X)|\geq t} (E - E(X))^2p(u) \\ V(X) &\geq \sum_{u\in\Omega||X-E(X)|\geq t} t^2p(u) = t^2P(\{|X - E(X)| \geq t\}) \end{aligned}$$

The final line yields the following famous result, called Chebyshev's inequality.

Theorem 27. *Let X be a discrete random variable such that $V(X)$ exists. Then*

$$P(\{|X - E(X)| \geq t\}) \leq \frac{V(X)}{t^2}$$

Notice that, as $t \to \infty$, then $P(\{|X - E(X)| \geq t\}) = 0$, by Chebyshev's inequality. This result confirms the intuition proffered by conceptualizing induced probability spaces as particle systems whose elements are the values of a random variable. In other words, the likelihood of the event a value of X is arbitrarily far from its expected value is negligible. We can strengthen this intuition by making a significant substitution.

Corollary 4.1. *Let X be a discrete random variable such that $V(X)$ exists and $c \in \mathbb{R}_{\geq 0}$. Then*

$$P(\{|X - E(X)| \geq c\sigma_X\}) \leq \frac{1}{c^2}$$

The corollary is obtained from Chebyshev's inequality by substituting $t = c\sigma_X$. Its relevance is that it measures the likelihood of a value of X being far from its expected value in terms of the standard deviation. In particular, it says that the likelihood of this event is bounded above by the inverse square of the number of standard deviations away from the expected value X is. Clearly, as $c \to \infty$, we have $P(\{|X - E(X)| \geq c\sigma_X\}) \to 0$.

4.8 Exercises

5 Parametric Families of Discrete Random Variables

In this chapter, we introduce the most famous parametric families of discrete random variables. Elements of these families are determined by the selection of parameter values, which in turn are themselves determined by the specifics of an experiment. These examples arise in the literature according to the ubiquity of their applications and are in many respects the extent to which the general public is familiar with probability theory altogether. We, however, can view at least a few of them as an introduction to stochastic processes which we will introduce in a generality appropriate to this textbook in chapter 6 in order to bridge the gap between the probability theory of this textbook and the algebraic interpretation of Markov processes commons in applications. The introduction to stochastic processes we have in mind are so-called Bernoulli processes, defined below.

5.1 Bernoulli Processes

In this section we introduce a discrete probability space that depends upon discrete values of time. This seemingly innocent hypothesis-that of a time dependence-shall introduce an entirely new class of probability spaces that we will rely upon to construct Markov chains in chapter 6. It is remarkable how they are already applicable to parametric families of random variables.

We begin by defining a Bernoulli trial to be a discrete probability space $(B, \beta = 2^B, P)$ where $B = \{s, f\}$ consists of two mutually exclusive simple outcomes, referred to informally as success and failure, respectively, and $P(s) = p \in [0, 1]$. The last hypothesis implies that $P(f) = 1-p = q$. It is trivial to show that a Bernoulli trial is a probability space. Indeed, this object is so simple, we denote it by B alone. Equally as elementary is the description of the random variable $X : B \to \mathbb{R}$ defined by $X(s) = 1$ and $X(f) = 0$. It is also straightforward to show the induced probability space is a discrete probability space itself. Bearing these definitions in mind, we next work to define the ambient probability space required for several of the parametric families below.

Definition 41. *We say a Bernoulli process is a sequence of Bernoulli trials $\mathscr{B} = \{B_i\}_{i=1}^{\infty}$ such that B_i is independent of B_j for any $i \neq j$. Furthermore, we insist that $P(s) = p$ and $P(f) = q$ for fixed p and all $i \geq 1$. Last, we say B_i is performed at time i.*

We shall adopt a somewhat informal approach to representing simple outcomes in a Bernoulli process. In the next chapter we shall undertake a more precise formulation. The reason to postpone a more accurate representation is to utilize the famous parametric families belwo as both motivations for and introductions to stochastic processes.

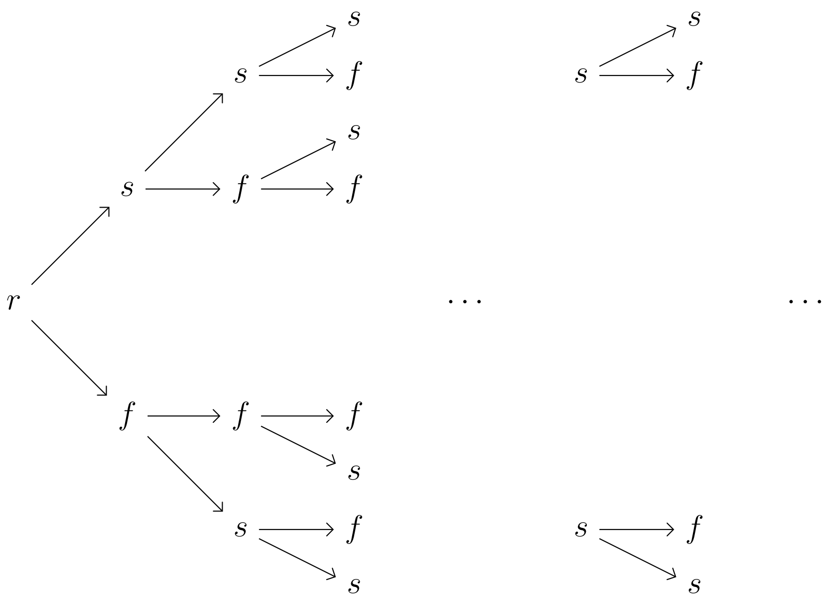
A simple event in a Bernoulli process is an arbitrary sequence of successes and failures, or s’s and f’s. Suppose we consider of a simple event up until time n, which means there are exactly n elements in the sequence representing it, say for example, $sfs \cdots fs$. Then the sequence in this example represents the event $s \cap f \cap s \cap \cdots f \cap s$, that is to say, ’$s$ occurs at time 1 and f occurs at time 2 and s occurs at time 3 and $\cdots$ and s occurs at time n.’ We extend this representation to arbitrary simple events in the obvious way. Now the probability of this simple event, in general, is given by

$$\begin{aligned} P(sfs \cdots s) =& P(s \cap f \cap s \cap \cdots f \cap s) \\ =& P(s)P(f \,|\, s)P(s \,|\, s \cap f) \cdots P(s \,|\, s \cap f \cap s \cap \cdots \cap f) \\ =& P(s)P(f)P(s) \cdots P(f)P(s) \end{aligned}$$

where the penultimate line is the general probability of the intersection of a sequence of events discussed in chapter 3 and the last line is the consequence of the hypothesis that the Bernoulli trials performed at each time value in the sequence are independent, that is, B_i and B_j are independent for $1 \leq i, j$ and $i \neq j$.

Next we introduce an important visualization of such a time dependent experiment. Specifically, we introduce *binary trees* which are graphs such that, at each height, there are exactly two edges attached to each vertex. The graph begins at a *root* which is a vertex with no predecessor. The branches in a binary tree, which are paths beginning at the root and containing exactly one edge for every vertex it passes through, are intended to represent the simple events in a Bernoulli process. Indeed, if we were to follow the approach of chapter 2 to describing such a process set theoretically, it would entail tuples in the Cartesian product of the elements of the Bernoulli process. To wit, $\Omega_{\mathscr{B}} = \times_{i=1}^{\infty} B_i$ is how we defined a process in chapter 2. We may, and should, think of the concatenation of labels for a sequence vertices connected by edges below as our representation of the corresponding tuple in the Cartesian product determined by the Bernoulli process.

Suppose $\mathscr{B} = \{B_i\}_{i=1}^{\infty}$ is a Bernoulli process, then we use the following binary tree to represent the set of simple events in $\Omega_{\mathscr{B}}$



so, for example, up until the n-th trial B_n following the $\cdots$ the uppermost branch represents the simple event $sss\cdots ss$. We shall denote this binary tree by $\Omega_{\mathscr{B}}$ which is the same notation as we use for a process in chapter 2 and write it in terms of its simple events by $\Omega_{\mathscr{B}} = \{a_1a_2\cdots a_i\cdots \mid (a_1, a_2, \ldots, a_i, \ldots) \in \times_{i=1}^{\infty} B_i \ , \ a_i = \{s \text{ OR } f\}_i \ , i \geq 1\}$.

Now we say a vertex s or f in this tree is *of height n* if the minimum number of edges connecting it to the root r is n. Equivalently, as well as interchangeably in the prose below, we shall say the same vertex *occurs at time n*. In the same straightforward manner, we shall refer to height n in the tree or time n in the tree to mean the set of edges whose tips are of height n. Notice that time n in the tree determines of a set of 2^{n-1} vertices whose ramifications or offshoot-edges determine the simple events of the Bernoulli trials that occur at time n. Let us label the vertices that occur at time n by $j = 1, 2, \ldots, 2^{n-1}$ and take $\Omega_n = \cup_{j=1}^{2^{n-1}} B_{n,j}$ where $B_{n,j} = \{s_j, f_j\}$ are the ramifications of the j-th vertex of height $n-1$. We say that Ω_n are the *outcomes at time n*. In our present notation, we can state and prove the following proposition.

Proposition 16. *Let $\mathscr{B}$ be a Bernoulli process and Ω_n be the outcomes at time n. Then $(\Omega_n, \beta = 2^{\Omega_n}, P)$ is a discrete probability space, where*

$$P(b_{n,j}) = \frac{\sum P_{n,j}(a_{n,j})}{2^{n-1}}$$

where $b_{n,j} = (a_{n,j})_{j\in J}$ and J is some discrete subinterval of $j = 1, 2, 3, \ldots, 2^{n-1}$.

As usual, we must show that P satisfies Kolmogorov's axioms in order to verify the assertion of the proposition. We could proceed by induction on the time variable n. However, we shall leave this proof to the reader, as it is routine.

In light of the proposition, we introduce an abuse of notation that is central to our eventual designs for processes more general than Bernoulli ones. Additionally, at the present time, we shall find applications for these concepts below when discussing parametric families of discrete random variables. Indeed, define the random variable $X_n : \Omega_n \to \mathbb{R}$ by the formula $X(b_{n,j}) = |b_{n,j}|_s$, where the notation $|b_{n,j}|_s$ means 'the number of successes s in $b_{n,j}$'. This in turn helps us to define the following object.

Definition 42. *Let $\mathscr{B}$ be a Bernoulli process and $X_n : \Omega_i \to \mathbb{R}$ such that $b_{i,j} \mapsto |b_{i,j}|_s$ be the number of successes at time $i \geq 1$ random variable. Then we say the sequence of discrete random variables*

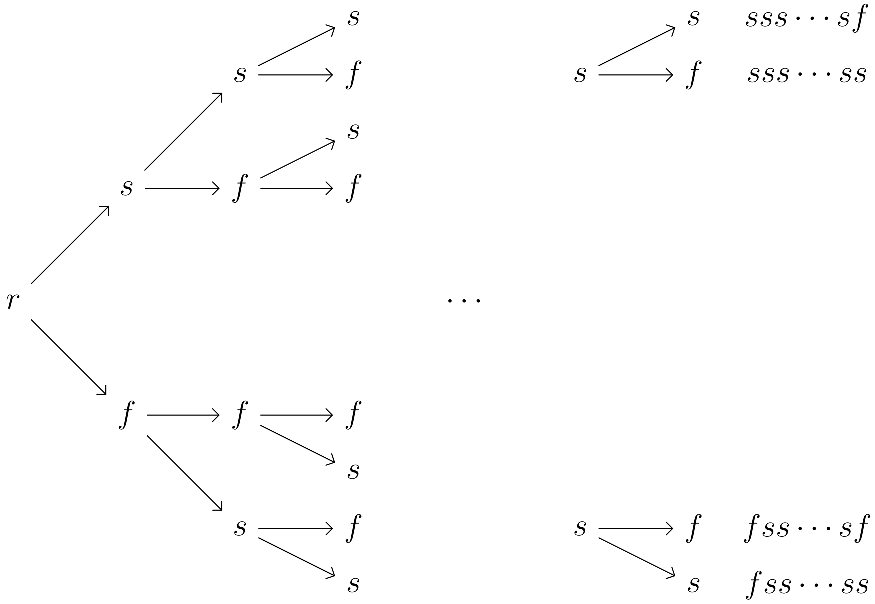
$$\{X_i\}_{i=1}^{\infty}$$

is a Bernoulli process.

I am not sure whether anymore examples than the main one are appropriate in this section, however, I should add an exercise for students to construct a similar example on their own.

5.2 The Binomial Random Variable $B(n,p)$

In this section we introduce the most well-known parametric family of discrete random variables, that of the *Binomial random variable*, denoted $B(n,p)$ below. To this end, let us consider a Bernoulli process of time n, that is, a sequence of Bernoulli trials $\mathscr{B}(n,p) = \{B_i\}_{i=1}^{n}$, where $p = P(s)$. In terms of $\Omega_{\mathscr{B}(n,p)}$ from the previous section, we may regard such a process as a binary tree of maximum height n.



where the right most column in the above diagram enumerates the sequences in $\Omega_{\mathscr{B}(n,p)}$ Accordingly we define the following discrete random variable $B(n,p) : \Omega_{\mathscr{B}(n,p)} \to \mathbb{R}$ such that $b \mapsto |b|_s$ where $|b|_s$ is the number of successes in $b \in \Omega_{\mathscr{B}(n,p)}$. We interpret this value as the number of successes in n Bernoulli trials. We have the following theorem.

Theorem 28. *Let $\mathscr{B}(n,p)$ be a Bernoulli process of time n and $(\Omega_{\mathscr{B}(n,p)}, \beta, P)$ the attendant discrete probability space. Define*

$$B(n,p) : \Omega_{\mathscr{B}(n,p)} \to \mathbb{R}$$

by $B(n,p)(b) = x$. Then $(\Omega_{B(n,p)}, \beta_{B(n,p)}, p_B(x))$ is a discrete probability space, where $0 \leq x = |b|_s \leq n$

Let us show that $p_B(x)$ satisfies Kolmogorov's axioms, as the first two datum already satisfy the requirements we have imposed upon such a triple for it to be a discrete probability space, in particular, we note that $\Omega_{\mathscr{B}(n,p)} = \{0, 1, 2, \ldots, n\}$. Let us first simply provide the formula for $p_B(x)$ by the multiplication rule and counting x-combinations. The process of computing $p_B(x)$ has two steps. The first is computing the probability of a simple event $b \in \Omega_{\mathscr{B}(n,p)}$ and the second is computing the size of the fibre of $B(n,p)$ over x. Therefore, consider b such that $x = |b|_s$ for $0 \leq x \leq n$. Then $P(b) = p^x q^{n-x}$ by the independence of the B_i. Now $|\{B(n,p) = x\}|$ is the number of branches b in $\Omega_{\mathscr{B}(n,p)}$ whose n edges are labeled by x successes s. In other words, $\binom{n}{x}$. Therefore, by the multiplication rule and the definition of $p_B(x)$ we have

$$p_B(x) = \tbinom{n}{x} p^x q^{n-x}$$

Second, let us turn to verifying this formula satisfies the axioms. Clearly $p_B(x) \geq 0$ for $0 \leq x \leq n$. Furthermore, $p_B((\Omega_{B(n,p)}) = 1$ since

$$\begin{aligned} p_B((\Omega_{B(n,p)}) =& p_B(0 \cup 1 \cup \cdots \cup n) \\ =& \sum_{x=0}^{n} p_B(x) \\ =& \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \\ =& (p+q)^n \\ =& 1 \end{aligned}$$

by the Binomial theorem of chapter 2. The third axiom of Kolmogorov's follows from the formalism of chapter 4 with respect to discrete subsets of $\Omega_{B(n,p)}$, concluding our proof.

Definition 43. *We say $B(n,p)$ is the Binomial random variable. We say that both n and p are its parameters, where n is the number of Bernoulli trials and p is the probability of success in the Bernoulli trials of a Bernoulli process of time n.*

It is more accurate to refer to $B(n,p)$ as a parametric family of discrete random variables, for there is one random variable corresponding to each pair of parameter values (n,p) satisfying the requirements. However, we abuse language by referring to $B(n,p)$ as the Binomial random variable to remain consistent with the broader literature.

Next we shall compute both the expected value and variance of $B(n,p)$ by the formal method in chapter 4. So, recall that $E(B(n,p)) = \sum_{x \in \Omega_{B(n,p)}} x p_B(x) = g'_B(1)$ where $g_B(z)$ is the generating function of $B(n,p)$. Observe,

$$\begin{aligned} g'_B(1) =& \\ \sum_{x \in \Omega_{B(n,p)}} x p_B(x) =& \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} \\ =& \sum_{x=1}^{n} n \binom{n-1}{x-1} p^x q^{n-x} \\ =& np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{(n-1)-(x-1)} \\ =& np \end{aligned}$$

where the last line follows from the binomial theorem.

The variance can also be computed formally. Recall that

$$V(B(n,p)) = g''_B(1) + g'_B(1) - (g'_B(1))^2$$

which is equal to $g_B''(1) + np - n^2p^2$ by the above argument. Let us finish the computation by evaluating $g_B''(1)$.

Observe that $g_B''(z) = \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^xq^{n-x}z^{x-2}$, so

$$\begin{aligned}
g_B''(1) =& \sum_{x=0}^{n} x^2\binom{n}{x}p^xq^{n-x} - np\\
=& \sum_{x=0}^{n} xn\binom{n-1}{x-1}p^xq^{n-x} - np\\
=& np\sum_{x=0}^{n} x\binom{n-1}{x-1}p^{x-1}q^{(n-1)-(x-1)} - np\\
=& np\sum_{y=0}^{m}(y+1)\binom{m}{y}p^yq^{m-y} - np\\
=& np\left(\sum_{y=0}^{m} y\binom{m}{y}p^yq^{m-y} + \sum_{y=0}^{m}\binom{m}{y}p^yq^{m-y}\right) - np\\
=& np\left(\sum_{y=0}^{m} m\binom{m-1}{y-1}p^yq^{m-y} + \sum_{y=0}^{m}\binom{m}{y}p^yq^{m-y}\right) - np\\
=& np\left((n-1)p\sum_{y=0}^{m}\binom{m-1}{y-1}p^{y-1}q^{(m-1)-(y-1)} + \sum_{y=0}^{m}\binom{m}{y}p^yq^{m-y}\right) - np\\
=& np\left((n-1)p+1\right) - np\\
=& (np)^2 + npq - np
\end{aligned}$$

Substituting our work into the formula for $V(B(n,p))$ given above, we have $V(B(n,p)) = (np)^2 + npq - np + np - (np)^2 = npq$, where $q = 1-p$. Collectively, these computations prove the following proposition that summarizes our work.

Proposition 17. *Let $B(n,p)$ be the Binomial random variable, then its generating function is given by*

$$g_B(z) = (pz+q)^n$$

and

1. $E(B(n,p)) = np$
2. $V(B(n,p)) = npq$

Notice in the statement of the proposition that we simplify the expression for the generating function. One can provide an alternative proof of the same now by differentiating with respect to z and applying the formulas of chapter 4.

Now do some examples. **EXAMPLES**
Example 5.3.1 Write some stuff *Q.E.F*
Example 5.3.2 Write some stuff *Q.E.F*
Example 5.3.3 Write some stuff *Q.E.F*

5.3 The Negative Binomial Random Variable $Q(r, p)$

In this section we consider the next parametric family of random variables, again predicated upon a Bernoulli process, that of the *Negative Binomial random* variable, denoted $Q(r, p)$ below. To prove its induced probability space satisfies Kolmogorov's axioms, we require the following theorem, known as the negative Binomial theorem. The Negative Binomial random variable $Q(r, p)$ is thus named for its reliance upon the conclusion of the following result. One demonstrates this result by computing the MacLaurin series expansion of $f(x) = (1 + x)^{-n}$.

Theorem 29. *Let n be a positive integer. Then*

$$\frac{1}{(1+x)^n} = \textstyle\sum_{k=0}^{\infty} \binom{n-1+k}{k}(-1)^k x^k$$

for $|x| < 1$

Let $\Omega_{\mathscr{B}}$ be a Bernoulli process and define the following discrete random variable $Q(t, r) : \Omega_{\mathscr{B}} \to \mathbb{R}$ such that $b \mapsto |b|_{rf}$ where $|b|_{rf}$ is the number of failures until the r-th success occurs in b. We interpret this value as the number of failures until the r-th success occurs. We have the following theorem.

Theorem 30. *Let $\mathscr{B}$ be a Bernoulli process and $(\Omega_{\mathscr{B}}, \beta, P)$ the attendant discrete probability space. Define*

$$Q(r, p) : \Omega_{\mathscr{B}} \to \mathbb{R}$$

by $Q(r, p)(b) = t$. Then $(\Omega_{Q(r,p)}, \beta_{Q(r,p)}, p_Q(t))$ is a discrete probability space, where $0 \leq t = |b|_{rf}$

Let us show that $p_Q(t)$ satisfies Kolmogorov's axioms, as the first two datum already satisfy the requirements we have imposed upon such a triple for it to be a discrete probability space. In particular, we note that $\Omega_{Q(r,p)} = \{0, 1, 2, \ldots\}$ is in bijective correspondence with $\mathbb{Z}_{\geq 0}$. Let us simply provide the formula for $p_Q(t)$ by the multiplication rule and counting t-unorderings. The process of computing $p_Q(t)$ has two steps. The first is computing the probability of a simple event $b \in \Omega_{\mathscr{B}}$ and the second, computing the size of the fibre of $Q(t, r)$ over t. Therefore, we consider b such that $t = |b|_{rf}$ for $0 \leq t$. Then $P(b) = p^r q^t$ by the independence of the B_i. Now $|\{Q(r, p) = t\}|$ is the number of t-unorderings of r successes so there are $\binom{r-1+t}{t}$. Therefore, by the multiplication rule and the definition of $p_Q(t)$, we have

$$p_Q(t) = \tbinom{r-1+t}{t} p^r q^t$$

Second, clearly $p_Q(t) \geq 0$ for $0 \leq t$. Furthermore, $p_B((\Omega_{B(n,p)}) = 1$ since

$$\begin{aligned} p_B((\Omega_{Q(t,r)}) =& p_Q(0 \cup 1 \cup \cdots \cup t \cup \cdots) \\ =& \sum_{t=0}^{\infty} p_Q(t) \\ =& \sum_{t=0}^{\infty} \binom{r-1+t}{t} p^r q^t \\ =& p^r \sum_{t=0}^{\infty} \binom{r-1+t}{t} (-1)^{2t} q^t \\ =& p^r (1-q)^{-r} \\ =& p^r p^{-r} \\ =& 1 \end{aligned}$$

by the Negative Binomial theorem stated above. The third axiom of Kolmogorov's follows from the formalism of chapter 4 with respect to discrete subsets of $\Omega_{Q(t,r)}$, concluding our proof.

Definition 44. *We say $Q(r,p)$ is the Negative Binomial random variable. We say both r and p are its parameters, where r is the number of successes and p is the probability of success in the Bernoulli trials of a Bernoulli process.*

Next we shall compute both the expected value and variance of $Q(t,r)$ by the formal method in chapter 4. So, recall that $E(Q(r,p)) = \sum_{t \in \Omega_{Q(r,p)}} t p_Q(t) = g'_Q(1)$ where $g_Q(z)$ is the generating function of $Q(r,p)$. Observe,

$$\begin{aligned} g'_Q(1) =& \\ \sum_{t \in \Omega_{Q(t,r)}} t p_Q(t) =& \sum_{t=0}^{\infty} t \binom{r-1+t}{t} p^r q^t \\ =& \sum_{t=1}^{\infty} r \binom{r-1+t}{t-1} p^r q^t \\ =& r \sum_{t=0}^{\infty} \binom{r+t}{t} p^{r-1} q^{t+1} \\ =& \frac{rq}{p} \sum_{t=0}^{\infty} \binom{r+1+t-1}{t} p^{r+1} q^t \\ =& \frac{rq}{p} \end{aligned}$$

where the penultimate line uses the above computation that shows $Q(t, r-1)$ satisfies Kolmogorov's axioms.

To compute the variance of $Q(r,p)$ we shall directly appeal to its generating function and its derivatives. Observe,

$$\begin{aligned} g_Q(z) =& \sum_{t=0}^{\infty} \binom{r-1+t}{t} p^r q^t z^t \\ =& \sum_{t=0}^{\infty} \binom{r-1+t}{t} p^r (qz)^t \\ =& p^r \sum_{t=0}^{\infty} \binom{r-1+t}{t} (qz)^t \\ =& p^r \frac{1}{(1-qz)^r} \\ =& \left(\frac{p}{1-qz}\right)^r \end{aligned}$$

It is straightforward to compute $g'_Q(z) = \dfrac{p^r rq}{(1-qz)^{r+1}}$ and $g''_Q(z) = \dfrac{p^r q^2 r(r+1)}{(1-qz)^{r+2}}$. Notice $g'_Q(1) = \dfrac{rq}{p}$. So, it follows from chapter 4 that

$$\begin{aligned} V(Q(t,r)) =& g''_Q(1) + g'_Q(1) - \left(g'_Q(1)\right)^2 \\ =& \frac{q^2 r(r+1)}{p^2} + \frac{rq}{p} - \frac{r^2q^2}{p^2} \\ =& \frac{q^2r^2 + q^2r + rpq - r^2q^2}{p^2} \\ =& \frac{rq(q+p)}{p^2} \\ =& \frac{rq}{p^2} \end{aligned}$$

since $(p+q) = 1$. We summarize our work in the following proposition.

Proposition 18. *Let $Q(r,p)$ be the Negative Binomial random variable, then its generating function is given by*

$$g_Q(z) = \left(\frac{p}{1-qz}\right)^r$$

and

1. $E(Q(t,r))) = \dfrac{rq}{p}$
2. $V(Q(t,r)) = \dfrac{rq}{p^2}$

It is well-known that there is an equivalent definition in terms of the number of trials until the r-th success. Let $Q'(r,p)(b) = |b|_{rs}$ denote this random variable. Then, abusing the notation for t, we have that the equivalence between these two discrete random variables is given by $Q(r,p)(t-r) = Q'(r,p)(t)$. The following corollaries to both the theorem and this proposition is to give their results in terms of the equivalent definition $Q'(r,p)$. In the corollary, we write $|b|_{rs}$ for the number of trials at which the r-th success occurs. As we mentioned above, $|b|_{rs} = |b|_{rf} + r$, denoted both values by t, again, as an abuse of notation.

Corollary 5.1. *Let $\mathscr{B}$ be a Bernoulli process and $(\Omega_{\mathscr{B}}, \beta, P)$ the attendant discrete probability space. Define*

$$Q'(r,p) : \Omega_{\mathscr{B}} \to \mathbb{R}$$

by $Q'(r,p)(b) = t$. Then $(\Omega_{Q'(r,p)}, \beta_{Q'(r,p)}, p_{Q'}(t))$ is a discrete probability space, where $0 \leq t = |b|_{rs}$

We will not verify Kolmogorov's axioms in this case, however, we emphasize that

$$p_{Q'}(t) = \tbinom{t-1}{r-1} p^r q^{t-r}$$

is both obtained by an analogous counting argument and still satisfies Kolmogorov's axioms. Naively, the induced probability function for $Q'(r,p)$ is obtained from that of $Q(r,p)$ by substituting $t = t - r$ into the latter variable's probability function. Indeed,

$$\begin{aligned}\binom{r-1+t}{t} &= \binom{r-1+(t-r)}{t-r} \\ &= \binom{t-1}{t-r} \\ &= \binom{t-1}{t-1-(r-1)} \\ &= \binom{t-1}{r-1}\end{aligned}$$

One notes that $\Omega_{Q'(r,p)} = \{r, r+1, \ldots \mid r \in \mathbb{Z}_{\geq 0}\}$. Next we have both the expected value and variance of Q' obtained as a corollary of the expected value and variance of Q.

Corollary 5.2. *Let $Q'(r,p)$ be the Negative Binomial random variable, then*

1. $E(Q'(r,p)) = \dfrac{r}{p}$
2. $V(Q'(r,p)) = \dfrac{rq}{p^2}$

The proof of this corollary is an exercise in chapter 4. Indeed, $Q'(y,r)$ is a transformation of $Q(t,r)$. Recall then that $E(aX+b) = aE(X) + b$ and that, for $Y = f(X)$, it is true also that $E(Y) = E(f(X))$. As such, for $Q'(r,p) = f(Q(r,p)) = Q(r,p) + r$, we have

$$\begin{aligned} E(Q'(r,p)) =& \frac{rq}{p} + r \\ =& \frac{rq + rp}{p} \\ =& \frac{r}{p} \end{aligned}$$

Similarly, since in general $V(aX+b) = a^2V(X)$ and Q' is a translation of Q by a constant, thereby $V(Q(r,p)) = V(Q'(r,p))$ is a true statement.

Now work out some examples. Now do some examples. **EXAMPLES**
Example 5.4.1 Write some stuff *Q.E.F*
Example 5.4.2 Write some stuff *Q.E.F*
Example 5.4.3 Write some stuff *Q.E.F*

5.4 The Geometric Random Variable $G(p)$

In this section, we consider a special case of the Negative Binomial random variable $Q'(r,p)$, that of the *Geometric random variable*. It is defined by the Negative Binomial random variable with $r = 1$, that is to say, it computes the probability that the first success occurs on the t-th trial. Indeed, we have the following definition.

Definition 45. *We say $G(p)$ is the Geometric Random variable. We p is its parameter, where p is the probability of success in each Bernoulli trial of a Bernoulli process.*

Theorem 31. *Let $\mathscr{B}$ be a Bernoulli process and $(\Omega_{\mathscr{B}}, \beta, P)$ the attendant discrete probability space. Define*

$$G(p) : \Omega_{\mathscr{B}} \to \mathbb{R}$$

by $G(p)(b) = |b|_{1s}$ is the number of trials until the first success s in b. Then $(\Omega_{G(p)}, \beta_{G(p)}, p_G(t))$ is a discrete probability space, where $1 \leq t = |b|_{1s}$

It is straightforward to show that the induced probability space satisfies Kolmogorov's axioms. As usual the first two pieces of data satisfy the definition of a discrete probability space, noting that $\Omega_{G(p)} = \{1, 2, 3, \ldots\}$. As for $p_G(t)$, we notice that the first $t-1$ trials are failures so that only the t-th trial is a success. Thus by the independence of Bernoulli trials, we have $p_G(t) = q^{t-1}p$. The fact the induced probability space satisfies Kolmogorov's axioms is tantamount to the convergence of the Geometric series and the hypothesis $|q| < 1$. Observe,

$$\begin{aligned} p_G(\Omega_{G(p)}) =& p_G(1 \cup 2 \cup \cdots \cup y \cup \cdots) \\ =& \sum_{t=1}^{\infty} q^{t-1}p \\ =& p\frac{1}{1-q} \\ =& 1 \end{aligned}$$

since $p = 1 - q$.

Computing the generating function also depends upon the sum of the Geometric series. Observe,

$$\begin{aligned} g_G(z) =& \sum_{t=1}^{\infty} q^{t-1} p z^t \\ =& p \sum_{t=1}^{\infty} (qz)^t \\ =& \frac{p}{(1-zq)} \end{aligned}$$

Computing the derivatives, we have $g'_G(z) = \dfrac{pq}{(1-zq)^2}$ and $g''_G(z) = \dfrac{2pq^2}{(1-zq)^3}$. Hence $g'_G(1) = \dfrac{q}{p} = \dfrac{1}{p} - 1$. Notice this is the same as the expected value of $Q(1,p)$ in the previous section. Applying the transformation to obtain the expected value of $Q'(1,p)$ gives us our result below. As for the variance of the Geometric random variable, we have

$$\begin{aligned} V(G(y)) =& \left(g''_G(1)\right)^2 + g'_G(1) - \left(g'_G(1)\right)^2 \\ =& \frac{2pq^2}{p^3} + \frac{q}{p} - \frac{q^2}{p^2} \\ =& \frac{pq^2 + p^2 q}{p^3} \\ =& \frac{q^2 + qp}{p^2} \\ =& \frac{q(q+p)}{p^2} \\ =& \frac{q}{p^2} \end{aligned}$$

We summarize our work with the generating function in the following proposition.

Proposition 19. *Let $Q'(r,p)$ be the Negative Binomial random variable and $Q'(1,p) = G(p)$ be the Geometric random variable, then its generating function is given by*

$$g_G(z) = \frac{p}{(1-zq)}$$

and

1. $E(G(p)) = \dfrac{1}{p}$
2. $V(G(p)) = \dfrac{q}{p^2}$

Before turning to examples,we make note of the following forgetfulness property of the Geometric random variable. Consider the following conditional probability

$$P(\{G(p) > a\} \mid \{G(p) > b\})$$

for $a, b \in \Omega_{G(p)}$ such that $a > b$. Then we have, by both independence and the definition of the induced probability function of a discrete random variable, that

$$\begin{aligned}
P(\{G(p) > a\} \mid \{G(p) > b\}) =& \frac{P(\{G(p) > a\} \ \cap \ \{G(p) > b\})}{P(\{G(p) > b\})} \\
=& \frac{P(\{G(p) > a\})}{P(\{G(p) > b\})} \\
=& \frac{q^a}{q^b} \\
=& q^{a-b} \\
=& P(\{G(p) > a - b\})
\end{aligned}$$

so that the final probability is interpreted to mean that the Geometric random variable forgets the conditional relationship between the events $\{G(p) > a\}$ and $\{G(p) > b\}$.

Now do some examples. **EXAMPLES**

Example 5.5.1 Write some stuff *Q.E.F*

Example 5.5.2 Write some stuff *Q.E.F*

Example 5.5.3 Write some stuff *Q.E.F*

5.5 The Hypergeometric Random Variable $H(M, N, k)$

In this section we consider a relative frequency interpretation of the Binomial random variable. In particular, the probability of success in this class of problems in computed by the relative frequency interpretation of probability. We think of a a branch $b(M, N)$ in a Bernoulli process of time N, say $\Omega_{\mathscr{B}(N,p)}$, as a sample in a Bernoulli process of time N. The Hypergeometric random variable then is exactly the probability model for the number of successes s in this sample.

We are implicitly contrasting the Hypergeometric variable with the Binomial in the way we are implying that the Binomial variable is somehow approximate. It is approximate in the sense that we have not adequately explained how we obtain the probability of success, namely p, in a Bernoulli trial. Without addressing this issue any further, we can agree that p is somehow approximate whereas $\dfrac{M}{N}$ is exact in what follows below. Indeed, it is enlightening to substitute $p = \dfrac{M}{N}$ to compare these variables induced probability functions and accompanying statistics.

Let us consider a branch $b(M, N) \in \Omega_{\mathscr{B}(N,p)}$ with the following features:

1. $b(M, N)$ consists of N edges
2. $b(M, N)$ has M edges labeled s

We can construct a discrete probability space predicated upon this branch together with a discrete random variable, the *Hypergeometric random variable*, denoted $H(M, N, k)$.

Theorem 32. *Let $\Omega^k_{b(M,N)}$ be the set of subsets of edges from $b(M,N)$ of size k and $(\Omega^k_{b(M,N)}, \beta, P)$ the attendant discrete probability space. Define*

$$H(M,N,k) : \Omega^k_{b(M,N)} \to \mathbb{R}$$

where $e \mapsto |e|_s$ is the number of edges labeled by s in the subset e. Then $(\Omega_{H(M,N,k)}, \beta_{H(M,N,k)}, p_H(x))$ is a discrete probability space for $0 \leq |e|_s = x \leq k$

As usual, we shall verify Kolmogorov's axioms and make note of the fact that $\Omega^k_{b(M,N)}$ consists of the subsets of size k of the set of N edges of $b(M,N)$. Accordingly, $|\Omega^k_{b(M,N)}| = \binom{N}{k}$. To compute $p_H(x)$ we shall compute the size of the fibre $H(M,N,k)$ over $0 \leq x \leq k$, that is, $|\{H(M,N,k)(e) = x\}|$. Observe that we can compute this by the multiplication rule, in two steps. First, there are $\binom{M}{x}$ subsets of size x of M s's within the set of edges of size N we may select from without replacement for e such that $H(M,N,k)(e) = x$. Second, there are $\binom{N-M}{k-x}$ subsets of size $k-x$ of $N-M$ edges labeled f within the set of edges of size N we may select from with replacement for the same e. Therefore, $|\{H(M,N,k)(e) = x\}|$ is $\binom{M}{x}\binom{N-M}{k-x}$ and so, by definition of the induced probability function, we have

$$p_H(x) = \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}}$$

To prove this probability function satisfies Kolmogorov's axiom, we require the following proposition.

Lemma 5.3. $\binom{N+M}{j} = \sum_{l=0}^{j} \binom{M}{j-l}\binom{N}{l}$

The proof depends upon the Binomial theorem. Observe,

$$\begin{aligned}
(x+y)^{N+M} &= \\
(x+y)^N(x+y)^M &= \left(\sum_{k=0}^{N} \binom{N}{k} x^{N-k}y^k\right)\left(\sum_{l=0}^{M} \binom{M}{l} x^{M-l}y^l\right) \\
&= \sum_{k=0}^{N}\sum_{l=0}^{M} \binom{N}{k}\binom{M}{l} x^{N+M-k-l}y^{k+l}
\end{aligned}$$

and so we have

$$\sum_{j=0}^{N+M} \binom{N+M}{j} x^{N+M-j}y^j = \sum_{k=0}^{N}\sum_{l=0}^{M} \binom{N}{k}\binom{M}{l} x^{N+M-k-l}y^{k+l}$$

Comparing coefficients in this equality, we see that $j = k+l$ and $l = j-k$. Further, since $k \geq 0$, we have $k \leq j$. Therefore

$$\sum_{j=0}^{N+M} \binom{N+M}{j} x^{N+M-j}y^j = \sum_{j=0}^{N+M} \left(\sum_{l=0}^{j} \binom{M}{j-l}\binom{N}{l}\right) x^{N+M-j}y^j$$

so it follows that

$$\binom{N+M}{j} = \sum_{l=0}^{j} \binom{M}{j-l}\binom{N}{l}$$

To finish our proof with the aid of this lemma, notice that x must satisfy both inequalities $x \leq k$ and $k - x \leq N - M$. Combining these inequalities gives $M - (N - k) \leq x \leq M$. We may translate this range to $0 \leq x \leq k$ so that $\Omega_{H(M,N,k)} = \{0, 1, 2, \ldots, k\}$. Therefore, by definition, we have

$$\begin{aligned} p_H(\Omega_{H(M,N,k)}) =& \sum_{x=0}^{k} \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \\ & \frac{\sum_{x=0}^{k} \binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \\ =& \frac{\binom{N}{k}}{\binom{N}{k}} \\ =& 1 \end{aligned}$$

which follows from the lemma, as desired.

Definition 46. *We say $H(M, N, k)$ is the Hypergeometric random variable. We say M, N and k are its parameters, where M is the number of successes in a sample of size k in a set of size N.*

Let us now compute the expected value and variance in terms of the generating function, $g_H(z) = \sum_{x=0}^{k} \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} z^x$. Observe,

$$\begin{aligned} g'_H(1) =& \sum_{x=0}^{k} \frac{x\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \\ =& \sum_{x=0}^{k} \frac{M\binom{M-1}{x-1}\binom{N-M}{k-x}}{\binom{N}{k}} \\ =& \frac{kM}{N} \sum_{x=0}^{k-1} \frac{\binom{M-1}{x-1}\binom{(N-1)-(M-1)}{k-x}}{\binom{N-1}{k-1}} \\ =& \frac{kM}{N} \end{aligned}$$

where the sum in the penultimate line simplifies to 1 and by Kolmogorov's axiom, as it is the Hypergeometric random variable for parameter values $H(M, N, k-1)$. Therefore, $E(H(M, N, k)) = \dfrac{kM}{N}$.

To compute the variance, a basic calculus computation gives

$$g''_H(z) = \textstyle\sum_{x=0}^{k} \dfrac{x(x-1)\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} z^{x-2}$$

So

$$
\begin{aligned}
g''_H(1) =& \sum_{x=0}^{k} \frac{x(x-1)\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \\
\\
=& \sum_{x=0}^{k} x(x-1)\binom{N-M}{k-x} \frac{M(M-1)(M-2)!}{x(x-1)(x-2)!(M-x)!} \frac{k(k-1)(k-2)!(N-k)!}{N(N-1)(N-2)!} \\
\\
=& \sum_{x=0}^{k} \frac{M(M-1)(k)(k-1)\binom{M-2}{x-2}\binom{N-M}{k-x}}{N(N-1)\binom{N-2}{k-2}} \\
\\
=& \frac{M(M-1)(k)(k-1)}{N(N-1)\binom{N-2}{k-2}} \sum_{x=0}^{k-2} \binom{M-2}{x}\binom{N-M}{k-2-x} \\
\\
=& \frac{M(M-1)(k)(k-1)\binom{N-2}{k-2}}{N(N-1)\binom{N-2}{k-2}} \\
\\
=& \frac{M(M-1)(k)(k-1)}{N(N-1)}
\end{aligned}
$$

At last we may apply the usual formula $V(H(M,N,k)) = g''_H(1) + g'_H(1) - (g'_H(1))^2$ which gives

$$
\begin{aligned}
V(H(M,N,k)) =& \frac{M(M-1)(k)(k-1)}{N(N-1)} + \frac{kM}{N} - \left(\frac{kM}{N}\right)^2 \\
=& \frac{kM}{N}\left(\frac{(M-1)(k-1)}{N-1} + 1 - \frac{kM}{N}\right) \\
=& \frac{kM}{N}\left(\frac{N(M-1)(k-1) + N(N-1) - kM(N-1)}{N(N-1)}\right) \\
=& \frac{kM}{N}\left(\frac{N^2 - Nk - NM + kM}{N(N-1)}\right) \\
=& \frac{kM}{N}\left(\frac{(N-M)(N-k)}{N(N-1)}\right)
\end{aligned}
$$

We summarize these laborious computations in the following proposition.

Proposition 20. *Let $H(M,N,k)$ be the Hypergeometric random variable, then*

1. $E(H(M,N,k)) = \dfrac{kM}{N}$

2. $V(H(M,N,k)) = \frac{kM}{N}\left(\frac{(N-M)(N-k)}{N(N-1)}\right)$

One might note that, in contradistinction to the earlier sections, we do not offer a closed form for $g_H(z)$ in this proposition. This is because there is not a closed form for the generating function of this variable. We refer the reader to the broader literature to explore the extent to which one can write down a more compact formula for the generating function of the Hypergeometric random variable.

Now do some examples. **EXAMPLES**
Example 5.6.1 Write some stuff *Q.E.F*
Example 5.6.2 Write some stuff *Q.E.F*
Example 5.6.3 Write some stuff *Q.E.F*

5.6 The Poisson Random Variable $P(\lambda)$

In this section we consider the limit as $n \to \infty$ of Binomial random variables $B(n,p)$ defined on an interval of real numbers of finite length. We start by considering a process related to a Bernoulli process, that of a Poisson structure. Translating the naive limit of Binomial variables in terms of random variables determined by the Poisson structure allows us to understand the limit we mean to consider. In this sense, a Poisson structure is the continuous extension of a Bernoulli process at time n.

An important application in this section is to approximate the Binomial random variable $B(n,p)$ by the so-called Poisson random variable, $P(\lambda)$, determined by the convergent limit, when *n is small compared to p* or, more concretely, with its parameter $\lambda = np$. As a variable in its own right, the Poisson variable $P(\lambda)$ is interpreted to *count* the number of occurrences of certain events in a given interval of finite length that occur, completely at random, at a rate of λ in that interval. Thus, its induced probability function $p_\lambda(x)$ shall compute the probability x of certain events occur.

To accomplish these aims, let us consider a drastic simplification of the general case that is available in the literature. First, we construct the space in which events that happen at a given rate, but completely at random, occur. To this end, fix a positive real number $l > 0$ and partition the interval $[0,l]$ into disjoint intervals of equal width $\delta = \frac{l}{2^j}$, viz.

$$[0,l] = \cup_{i=1}^{n}(x_{i-1}, x_i] \cup \{0\}$$

for fixed $j \in \mathbb{Z}$. We regard each cell in this partition as the i-th episode of a process which transpires over a duration δ. Assume at each episode a Bernoulli trial B_i is performed and that $P(s) = p$ for each performance. Moreover, assume p is proportional to the duration of an episode, that is, $p = \lambda \cdot \delta$ for some fixed $\lambda \in \mathbb{R}$ such that $0 \leq p \leq 1$. Notice this assumption implies that $q = 1 - \lambda\delta$. We shall say that a certain event occurs at the i-th episode if the performance of B_i results in s, and likewise we shall say that a certain event does not occur at the i-th episode if B_i is performed and results in f.

Under these auspices, the interval $[0,l]$ is our model for the occurrence of certain events occurring at n episodes, completely at random, and at a particular, constant rate λ. We should remark that this constant rate in applications is taken to be the average rate of occurrence of a certain event. We say the interval has a *Poisson structure*.

Now we are able to achieve our stated aim by quantifying the occurrences of certain events by counting the number of their occurrences in a Poisson structure. To this end, we construct an $\mathbb{N}$-valued counting function $N_j : [0, l] \to \mathbb{R}$ on the Poisson structure, defined by the formula

$$N_j(t) = x \in \mathbb{N}$$

for $t \in (0, l]$ to be the number of occurrences of a certain event in $(0, t] \cap [0, l]$. Moreover, we insist that $N(0) = 0$ since, at time 0, no Bernoulli trial has yet been performed. Observe that $N_j(t)$ is a discrete random variable such that $\Omega_{N_j} = \{0, 1, 2, \ldots, n\}$ and whose induced probability function is given by

$$p_{N_j}(x) = \binom{n}{x} \left(\frac{\lambda}{2^j}\right)^x \left(1 - \frac{\lambda}{2^j}\right)^{n-x}$$

Let us pause to recall our stated intention at the outset of this section, which was to consider the limit of Binomial random variables. In the present context, we stress that such a limit arises as $\delta \to 0$, for this is tantamount to $j \to \infty$ or, equivalently, $n \to \infty$, since $\delta = \dfrac{l}{2^j}$. Before we proceed to pursue our objective through this observation, consider the following argument.

Let $B(n, p)$ be an arbitrary Binomial random variables. By Chebyshev's inequality, we have

$$P(\{|B(n, p) - np| \geq t\}) \leq \frac{npq}{t^2}$$

or, by the law of the unconscious statistician,

$$P(\{|\frac{B(n, p)}{n} - p| \geq t\}) \leq \frac{pq}{nt^2}$$

which tends to 0 as $n \to \infty$, implying that for arbitrary p, the sequence $\{B(n, p)\} \to 0$. Accordingly, the probability of x successes in n Bernoulli trials as n becomes arbitrarily large vanishes.

In order to obviate this conclusion from Chebyshev's inequality, we insist that *p is small compared to n*. In concrete terms, we insist that p is independent of n. This means that $np = \lambda$ for λ a fixed constant. Accordingly, by Chebyshev's inequality, we have

$$P(\{|B(n, p) - \lambda| \geq t\}) \leq \frac{\lambda q}{t^2}$$

which is now independent of n. Thus, the sequence $\{B(n, p)\}_{n=1}^{\infty}$ does not vanish as $n \to \infty$. We shall incorporate this hypothesis below so that the limit we obtain is a meaningful one in terms of the corresponding limit of Binomial random variables. Now, to continue from where we left off, let us state and prove the main theorem of this section.

Theorem 33. *Let $(\Omega_{N_j(t)}, \beta, p_{N_j}(x))$ be the induced probability space with respect to $N_j(t)$. Define*

$$P(\lambda) : [0,l] \to \mathbb{R}$$

where $P(\lambda)(t) = x$ is the number of occurrences of a certain event in $(0,t] \cap [0,l]$, where $[0,l]$ is equipped with a Poisson structure. Then $\lim_{j\to\infty} N_j(t) = P(\lambda)$ exists non-trivially if p is small compared to n and $(\Omega_{P(\lambda)}, \beta, p_\lambda(x))$ is a discrete probability space.

We shall verify that $p_\lambda(x) = \lim_{j\to\infty} p_{N_j(x)}$ when p is small compared to n, that is, $\lambda = np$. We emphasize this computation is equivalent to $p_\lambda(x) = \lim_{n\to\infty} p_{B(n,p)}(x)$ since n is the length of the partition of the Poisson structure. Indeed, we use this equality below. Observe,

$$\begin{aligned}
\lim_{j\to\infty} p_{N_j}(x) &= \lim_{j\to\infty} \binom{n}{x} p^x q^{n-x} \\
&= \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{2^j}\right)^x \left(1 - \frac{\lambda}{2^j}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(\frac{n!}{(n-x)n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n q^{-x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(\frac{n!}{(n-x)n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} e^{-\lambda}
\end{aligned}$$

so we conclude that $p_\lambda(x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$. Next we show that this probability function satisfies Kolmogorov's axioms.

As usual, the first and third axioms are left to the reader. Notice that as $\delta \to 0$ we have $n \to \infty$, so that $\Omega_{P(\lambda)} = \{0, 1, 2, 3, \ldots\}$. So, to show the second axiom, we have

$$\begin{aligned}
p_\lambda(\Omega_{P(\lambda)}) &= \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\
&= e^{-\lambda} e^{\lambda} \\
&= 1
\end{aligned}$$

as desired.

Definition 47. *We say $P(\lambda)$ is the Poisson random variable. We say that λ is its parameter where λ is the average number of successes in a Poisson process.*

Next, let us compute the moment generating function in order to obtain the expected value and variance of the Poisson random variable. Let us compute both $g\prime_\lambda(1)$ and $g''_\lambda(1) +$

$g'\lambda(1) - (g'_\lambda(1))^2$. Applying the definition, we have

$$\begin{aligned} g_\lambda(z) =& \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} z^x \\ =& e^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} \\ =& e^{-\lambda} e^{z\lambda} \\ =& e^{\lambda(z-1)} \end{aligned}$$

which is a rather convenient closed form for the Poisson variable's generating function. It is trivial that $g'_\lambda(1) = \lambda$ and $g''_\lambda(1) = \lambda^2$ and the above formulas simplify. We summarize our work in the result of the following proposition.

Proposition 21. *Let $P(\lambda)$ be the Poisson random variable. Then its generating function is given by*

$$g_\lambda(z) = e^{\lambda(z-1)}$$

and

1. $E(P(\lambda)) = \lambda$
2. $V(P(\lambda)) = \lambda$

In the above discussion, we have computed the probability of a certain number of events in an interval of finite length occurring at a rate of λ. A good question we can answer is how does this figure change if only the length of the interval changes. In other words, since the probability is parameterized by the rate of occurrence λ, how does λ change? The answer is given by the following formula.

Lemma 5.4. *Let $P(\lambda)$ be the Poisson random variable on an interval of length l. Then P extends to an interval of length $k = hl$ by the formula $P(h\lambda)$.*

Before proceeding to examples, we provide a corollary to the above theorem demonstrating that $P(\lambda)$ is a discrete random variable. Regarding the work we invested in that demonstration, we see that the Poisson random variable approximates the Binomial random variable in general when the parameter of the latter satisfy the hypothesis that p is small compared to n. As a rule, one interprets this statement to mean precisely that $n \geq 100$ and $np \leq 10$. Accordingly, we have the following corollary.

Corollary 5.5. *Let $P(\lambda)$ be the Poisson random variable and $B(n, p)$ the Binomial random variable. Suppose that p is small compared to p. Then $P(\lambda) \cong B(n, p)$ with $\lambda = np$.*

Now do some examples. **EXAMPLES**
Example 5.7.1 Write some stuff *Q.E.F*
Example 5.7.2 Write some stuff *Q.E.F*
Example 5.7.3 Write some stuff *Q.E.F*

5.7 The Multinomial Random Variable

In this section we consider the joint distribution that generalizes the Binomial random variable at the beginning of this chapter.

Definition 48. *Let $(X_1, X_2, \ldots, X_n)$ be an n-tuple of discrete random variables and $Z = f(X_1, X_2, \ldots, X_n)$ then is a Discrete*

Let us first show that the probability function is give by

$$p(x_1, x_2, \ldots, x_n) = \tbinom{n}{x_1!x_2!\cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}$$

5.8 Exercises

The exercise format should follow a pattern alike that of the exercises. So we should have

Exercise 5.2.1 For an exercise from 5.2

Exercise 5.3.1 For an exercise from 5.3

6 Markov Chains

This chapter introduces one of the main mathematical objects used in applications in the modern world, that of the Markov chain. In this textbook a Markov chain is a triple of linear algebraic data satisfying a single recurrence relation. The nature of this object may at first mystify a student for its apparent lack of relationship to probability theory, specifically, the formalism of discrete probability spaces and random variables. Pursuant to this concern, we first introduce graphs and stochastic processes in a manner whose rigor is appropriate for undergraduates to establish the relationship between the content of this text so far and a Markov chain. This material was foreshadowed by the informal introduction in the previous chapter to Bernoulli processes. We hope the frequent use of processes in the previous chapter convinces the reader there is more to discover!

The intuition of binary trees and sequences of random variables in chapter 5 extends in the straightforward manner to arbitrary trees and sequences of random variables in the present one. We call such extensions stochastic processes, below. Stochastic processes, together with a significant axiomatic simplification of the probabilities of simple events become Markov *processes*, which are generalizations of Bernoulli processes in chapter 5. However, as they are more general, we interpret the axiom distinguishing these processes in terms of linear algebra by way of the law of total probability discussed in chapter 1, rather than through the counting arguments that simplified Bernoulli processes. Once we have this, we endeavor to translate all of the data contained within a Markov process into linear algebra to arrive at the Markov chain.

After we translate Markov processes into linear algebra as Markov chains, we abandon the explicit use of probability theory for the remainder of the chapter and in the following one on Markov Decision Processes. It is our hope that this most important chapter convinces the reader why a modern student of computer science would be interested in the foundations of discrete probability theory.

6.1 Introduction to Graphs and Trees

In this section we introduce graphs and trees together with their matrix representations. These concepts are necessary for several purposes relevant to Markov chains. Initially, however, we use them to make sense of an experiment that occurs in stages. It is from here that we are able to transfigure the formalism of random variables into linear algebraic data, which will simplify matters remarkably.

Definition 49. *A graph G is a pair of sets V and E, referred to as vertices and edges, respectively, denoted $G = (V, E)$. The set V is an ordinary set whose elements are referred to as vertices. Furthermore, E is the set of all singletons and pairs of elements in V whose elements in either case are referred to as edges. We say G is a finite graph if $|V|$ is finite.*

Identifying an edge with the elements of V comprising it is in fact a function, $\epsilon : E \to 2^V$ such that $e \mapsto \{v, v'\}$ under ϵ, where v or v' may be the empty set, but not both. We say the image of e under ϵ are its *endpoints.* What follows are a few definitions of terms one requires to speak cogently upon matters of graphs and their applications.

We say an edge with a single endpoint is a *loop.* Two distinct edges with the same endpoints are *parallel.* Two vertices connected by an edge are said to be *adjacent.* In particular, we adopt this definition to loops in the obvious way. A vertex with no edges, that is to say, the fibre over $\{v\} \in 2^V$ with respect to ϵ is empty, is *isolated.* Lastly, we say an edge e is *directed* if its endpoints are an ordered set. We say a graph is *directed* if each of its edges are directed.

As is usual in mathematics, we may consider substructures of a given one, so that we define a *subgraph* of G to be a pair of subsets $H = (V', E')$ of G such that $\epsilon(e') \in \epsilon(E)$ for each $e' \in E'$. More plainly, we require that all of the endpoints of edges in H are in the endpoints of the edges of G.

Next, although brief, we have the important concept of degree. Specifically, the degree of a vertex $v \in V$ is the number of $e \in E$ such that $v \in \epsilon(e)$. Intuitively, the *degree* of a vertex is the number of edges that contain it. An interesting fact of graphs is that the sum of the degrees of all its vertices is equal to twice the number of its edges. We shall not make use of this fact, as the purposes of graphs in this text is to either model time-dependent experiments or to determine Markov chains via their matrix representations. Pursuant to these motivations, again, what follows are a few necessary definitions.

Let $v, v' \in V$, then a *walk from v to v'* is a finite sequence of edges, say $e_1, \ldots, e_k$, such that $v \in \epsilon(e_1)$ and $v' \in \epsilon(e_k)$ and $e_i \cap e_j \neq \emptyset$ for $1 \leq i, j \leq k$. Notice this definition does not exclude parallel edges, so to stress this feature, we say a *path* is a walk with no parallel edges. Next, neither the definition of walk nor the definition of path exclude repeated vertices, so a path without repeated vertices is called a *simple path.* Lastly, a *closed walk* is a walk that begins and ends at the same vertex, and a *circuit* is a closed walk that is also a path. A *simple circuit* is a circuit and a simple path. Finally, before stating the definition of a tree, we define a graph G to be *connected* if for any two vertices there exists a walk from one to the other.

Definition 50. *Let T be a graph. We say T is a tree if T has no non-trivial circuits and is connected. Assume that T is a tree with at least three vertices, then a vertex of degree one is* *called a leaf and a vertex of degree greater than one is called a branch vertex. We write L_T for the set of leaves in T.*

One can refine the structure of trees by distinguishing a vertex of degree one as a *root*. In this way a root is emphatically not a leaf. We usually omit the root from the graphical representation of a tree and instead denote its offshoot or ramification by r. Moreover, with respect to such a declaration, one can then define the *height* of a vertex v to be the least number of edges required to connect v to r, denoted $ht(v)$. In applications, we shall use the height of a vertex to reflect the episode in a time dependent experiment at which a subexperiment occurs. If we declare a root of T, then we say T is a *rooted tree.* Given a vertex v, we say the *ramifications* of v are the vertices $v' \in V_T$ such there exists an edge $e \in E_T$ containing both v and v' as endpoints and that the height $ht(v') = ht(v) + 1$. The simple paths connecting the root r to a leaf $l \in V$ are called *branches*, and lastly, the height of the leaves in T determine the *height of* T. Of course, the height of T is the maximal length of the branches. A *subbranch* is a branch whose endpoint is a branch vertex.

6.2 Matrices and Matrix Representations of Graphs

In this section we review matrices and indicate how they are related to graphs as we introduced them in the previous section. In particular, we turn our attention to matrix representations of graphs. Associating a certain matrix to a graph G will be a major aspect of computing the probability a Markov chain resides in a particular state at a particular time, below. We begin with a short introduction to matrices, their sums, and products. We refer the reader to the literature for a more careful treatment.

Skip this stuff for this semester. Define an $m \times n$ matrix M. Define the sum of two rectangular matrices M and N by adding their entries. Define the formula for the multiplication of matrices by the matrix whose (i, j) entry is the dot product of the i-th row and j-th column for matrices whose dimensions satisfying the relation $m \times n$ and $n \times p$ to get a $m \times p$ matrix. Stress repetition of the multiplication formula in order to compute higher powers of a square matrix.

Let $\Gamma = \{v_{ij}\}_{1 \leq i \leq m, 1 \leq j \leq n}$ be an $m \times n$-matrix and $G = (V, E)$ be a finite graph, then we say a function $\gamma : \epsilon(E) \to \mathbb{R}$ is a *matrix representation* of G if

$$\Gamma = \{v_{ij} = \gamma(\{v_j, v_i\})\}_{1 \leq i \leq m, 1 \leq j \leq n}$$

Observe that the formula for the (i, j)-entry of the matrix representation γ is determined by the edge joining the vertex v_j to the vertex v_i.

The first example we consider is that of the *adjacency matrix representation* of a graph G. To compute this we must compute the image of γ according to some rule obtained from the graph G, itself. Recall from the previous section that two vertices v_j and v_i are said to be adjacent if there exists $e \in E$ such that $\epsilon(e) = \{v_j, v_i\}$. For a fixed v_j, define the integer $adj(v_j) = |\{v_i \in V \mid \epsilon(e) = \{v_j, v_i\}\}|$. That is, the number of vertices in G adjacent to v_j. If v_j is an isolated vertex, we define $adj(v_j) = 1$. Then $adj(v_j)$ determines a matrix representation of G by the formula $\gamma(\{v_j, v_i\}) = adj(v_j)$ if there exists an edge $e \in E$ such that $\epsilon(e) = \{v_j, v_i\}$ and 0 otherwise. Notice, in general, that $adj(v_j) \neq adj(v_i)$. Let $\Gamma = [g_1 \ \cdots \ g_j \ \cdots \ g_n]$ to be the $m \times n$ matrix such that its j-th column is determined by the matrix representation γ

$$g_j = \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_m \end{bmatrix}$$

where $a_i = adj(v_j)$ if there exists $e \in E$ such that $\epsilon(e) = \{v_j, v_i\}$ and 0 otherwise, for $1 \leq i \leq m, 1 \leq j \leq n$. Notice that if $v_j \in V$ is isolated then the (j, j) entry of Γ is 1 and the other entries of g_j are 0. In terms of linear algebra, one says that the column g_j representing an isolated vertex is an elementary column. The presence of elementary columns shall be an obstruction we must study below.

6.3 Stochastic Processes

In this section we set-up for the important link between the discrete probability theory we have developed so far in terms of probability spaces and random variables and the formalism of Markov chains. Specifically, we introduce stochastic processes, which are specific subprocesses of processes in the sense of Chapter 2, for this purpose alone. We take Markov processes as special cases thereof and then we interpret these processes in the context of standard linear algebra in terms of the law of total probability. It is according to this interpretation that proffers the so-called Markov chains.

We begin by considering how to instantiate experiments that occur in n stages or, equivalently, are dependent upon n passing episodes of discrete time, as discrete probability spaces. To do this, we must begin by imposing some restrictions on the scope of such experiments. The following three hypotheses are derived from such necessary restrictions.

1. All time dependent experiments occur in n episodes of discrete time, where n is a finite, non-negative integer.

2. At any episode in an experiment satisfying the first hypothesis, we allow a finite number of subexperiments to be performed in which only finitely many simple outcomes may occur.

3. Supposing an experiment satisfies the first two hypotheses, we assume further that its simple events are determined by a discrete-time sequential sequence of events that occur in the subexperiments and whose outcomes are known.

We say any experiment satisfying these three hypotheses is a *stochastic process*. Moreover, we refer to these hypotheses collectively as the *stochastic axioms*. The third hypothesis is perhaps the most important, for we use it to help determine the likelihood of events in a stochastic process. The last hypothesis entails that the probability of an event in the entire experiment is determined by the probabilities of events in the sequences of subexperiments by which it is determined. The constructions below rely heavily upon this assertion.

Next is to represent stochastic processes as rooted trees via the following construction. Let T be a rooted tree of height n, $v \in V_T \setminus L_T$ a branch vertex of height i. We define for each such branch vertex its *ramification space* to be the set of subbranches in T whose endpoints are the ramifications of v. Let us denote the set of such subbranches by

$$\Omega_v(i) = \{b \in T \ \mid \ b = e_1 e_2 \cdots e_i \cdot e_{i+1}, \ \ , \epsilon(e_{i+1}) = \{v, v'\} \ \}$$

Please remember that our subbranches begin at the root in our notation, so that $\epsilon(e_1) = \{r, v\}$. Furthermore, notice that since v is not a leaf, this restricts the range of the height to $0 \leq i \leq n-1$.

We shall interpret these subbranches b in terms of a stochastic process. Given an arbitrary edge in b, say e_j for $1 \leq j \leq i$, such that $\epsilon(e_j) = \{v_j, v_{j+1}\}$, we interpret v_j to be the outcome of the subexperiment performed at time episode $j-1$ and v_{j+1} to be the outcome of the subexerpiment performed at time episode j given that v_j has occurred. Accordingly, each edge represents a conditional event in the stochastic process. We shall adopt the convention in examples that we equate vertex labels with the outcomes of subexperiments that they represent. This point is especially important to keep in mind below when we introduce outcome functions.

Now, more generally, we interpret an element $b \in \Omega_v(i)$ as an outcome in a stochastic process in the following manner. We decompose b into a sequence of conditional events by interpreting the edges that compose b in the above manner that associates conditional events to edges. This decomposition implies that b is interpreted as the simple outcome v' is the outcome of the subexperiment performed at time episode i given that v_1 and v_2 and $\cdots$ and v_i have occurred, or symbolically, $v' \,| v_1 \cap v_2 \cap \cdots \cap v$. This interpretation of a simple event in the ramification space of a vertex v is consistent with stochastic axiom 3. Given an interpretation of subbranches as simple events in a stochastic process we are ready now to associate it to a discrete probability space and thus fulfill the objective of this section.

Clearly the ramification space $\Omega_v(i)$ as described above is a finite set, so suppose that $|\Omega_v(i)| = k$. It follows that $\beta_v(i) = 2^{\Omega_v(i)}$ is finite as well. Following chapter one, we can define a probability function P_v^i with respect to a ramification space by the formula

$$P_v^i(A) = \sum_{l \,|\, b_l \in A} p_l$$

for $A \in \beta_v(i)$ by assigning numbers p_l to its simple events, say b_l, for $1 \leq l \leq k$, such that $\sum_{l=1}^k p_l^i = 1$. Our conclusion will be that $(\Omega_v(i), \beta_v(i), P_v^i)$ is a discrete probability space, as desired. Indeed, stochastic axiom 3 means P_v^i satisfies Kolmogorov's axioms. Still, it does not define its value on a simple event specifically, so we do that next.

Explicitly, define the probability function P_v^i using our formula for the probability of an intersection of event in terms of their conditional probabilities from chapter 3, as follows

$$\begin{aligned} P_v^i(v_1 \cap v_2 \cap \ldots \cap v_{i-1} \cap v \cap v_l') =& P_v^0(e_1) P_v^1(e_1 e_2) \cdots P_v^i(e_1 e_2 \cdots e_i) \\ =& P_v^0(v_1) P_v^1(v_2|v_1) P_v^2(v_3|v_1 \cap v_2) \cdots P_v^i(v_l' \,| v_1 \cap v_2 \cap v_3 \cap \cdots \cap v) \end{aligned}$$

As the simple event b_l is represented by $v_l' \,| v_1 \cap v_2 \cap v_3 \cap \cdots \cap v$, we can solve for the desired quantity $P_v^i(v' \,| v_1 \cap v_2 \cap v_3 \cap \cdots \cap v)$ which then yields the values p_l for our formula $P_v^i(A)$, viz.

$$\begin{aligned} P_v^i(b_l) =& \frac{P_v^i(v_1 \cap v_2 \cap \ldots \cap v_{i-1} \cap v \cap v')}{P_v^0(v_1) P_v^1(v_2|v_1) P_v^2(v_3|v_1 \cap v_2) \cdots P_v^i(v \,| v_1 \cap v_2 \cap v_3 \cap \cdots \cap v_{i-1})} \\ =& p_l^i \end{aligned}$$

Again we emphasize that it is stochastic axiom 3 that implies $\sum_{l=1}^{k} p_l = 1$. We can apply the above formula for $P_v^i(A)$ to any $A \in \beta_v(i)$, thus rendering at least the ramification spaces in terms of discrete probability theory. The aspect of a stochastic process reflected by the set of all ramification spaces is stochastic axiom 2.

Observe that we did not fix a height for all branches in a rooted tree T nor the stochastic process it represents. Indeed, we defined the height as the maximal length of a branch. It is therefore possible that some branches are shorter than others. In this case, stochastic axiom 3 implies how to interpret such situations. Indeed, suppose $P_v^i(b_l) = \alpha$, then $P_v^j(b_l) = \alpha$ for $j > i$. This compatibility condition between different episodes is the precise manner in which stochastic axiom 3 clarifies how to interpret branches of different lengths in a rooted tree representation of a stochastic process.

As we have defined the ramification spaces attached to a stochastic process in order to encode stochastic axiom 2, let us now introduce the instantiation of the stochastic process itself as a discrete probability space. Namely, let us construct the *branch space* associated to a stochastic process.

The branch space imparts the structure of a discrete probability space on the set Ω_T of branches in the tree T. Indeed, in terms of the ramification spaces defined above, we have

$$\Omega_T = \bigcup_{v \ | ht(v) = n-1} \Omega_v(n-1)$$

so that simple events in a stochastic process are represented by branches b in the tree T. As before, plainly Ω_T is a finite set and $\beta_T = 2^{\Omega_T}$ is, too. The following theorem defines a probability function which, together with the aforementioned data, assembles to a discrete probability space that is our discrete probability theoretic model of the stochastic process.

Theorem 34. *Let T be a rooted tree of height n and Ω_T its set of branches. Given any branch $b = e_1 e_2 \cdots e_n$, define*

$$P_B(b) = P(\cap_{i=1}^{n} v_i)$$

Then $(\Omega_T, \beta_T = 2^{\Omega_t}, P_B)$ is a discrete probability space.

We refer to the probability space of this theorem as the branch space of the stochastic process. The proof of this theorem is a consequence of the law of total probability in conjunction with the stochastic process hypothesis, that is, we assume that the probability of the events in the experiment at the i-th episode are known given knowledge of the probabilities of the previous $i-1$ episodes. Given a branch $b = e_1 e_2 \cdots e_n$, we can define its probability in terms of the conditional probability of the edges contained in the same. To wit,

$$\begin{aligned} P_B(b) =& P_B(e_1 e_2 \cdots e_n) \\ =& P_B(\cap_{i=1}^{n} v_i) \\ =& P_B(v_1) P_B(v_2 | v_1) P_B(v_3 | v_1 \cap v_2) \cdots P_B(v_n | v_1 \cap v_2 \cap v_3 \cap \ldots \cap v_{n-1}) \end{aligned}$$

Now to verify Kolmogorov's second axiom, we proceed by the law of total probability. Let us partition Ω_T by the so-called beginning partition. That is, we define cells of a partition according to the first vertex of the branch different from the root, that is, by the ramifications of the root itself. So, let $v_t \in \Omega_r(0)$ for $1 \leq t \leq u$, then write $\{v_t\}$ for the set of branches $b = e_1 e_2 \cdots e_n$ such that $\epsilon(e_1) = v_t$. Then, summing over t gives

$$P_B(\Omega_T) = \sum_{t=1}^{u} P_B(\{v_t\})$$

We want to show that $P_B(\{v_t\}) = P_B(v_t)$ for then, by our assumption on ramification spaces, we have

$$P(\Omega_T) = \sum_{t=1}^{u} P_r(v_t) = 1$$

as desired.

So, we have

$$\begin{aligned} P_B(\{v_t\}) =& P_B(\bigcup_{b\in\Omega_T \ |\epsilon(e_1)=v_t} b) \\ =& \sum_{b \ | \ \epsilon(e_n)=v} P_v^{n-1}(b) \\ =& \sum_{b \ | \ \epsilon(e_n)=v} P_v^{n-1}(v' \, |v_t \cap v_2 \cap v_3 \cap \cdots \cap v_{n-1}) \end{aligned}$$

Now repeatedly factoring out $P_r(v_t)$ from decreasing conditional events and applying the hypothesis that each ramification space satisfies Kolmogorov's axioms gives

$$P_B(\{v_t\}) = P_r(v_t)$$

as desired. Notice we assumed Kolmogorov's third axiom, or at least we invoke chapter one's result to justify the assumption. Moreover, the verification of the first axiom is obvious by the definition.

Having fulfilled our objective of rendering a stochastic process as a discrete probability space, we turn at last to the standard definition of a finite stochastic process in terms of a sequence of random variables. Define outcome spaces to be the union of vertices of a fixed height, as follows

$$\Omega_T(i) = \bigcup_{v \ | \ ht(v)=i} v$$

Now, for each i, we impose a total order on the i-th outcome space, denoted $v(1_i) < v(2_i) < v(3_i) < \cdots < v(w'_i)$, where $|\Omega_T(i)| = w'$. Given these numerical labels induced by the total order we can define a sequence of random variables in terms of *outcome functions* $O_i : \Omega_T \to \Omega_T(i)$, for $1 \leq i \leq n$ by $O_i(b) = \epsilon(e_i) \cap \Omega_T(i)$. Indeed, consider the random variables $Y_i : \Omega_T(i) \to \mathbb{R}$ for each such i, where $Y_i(v(j_i)) = j_i$. Then the we say the sequence of transformations $\{X_i\}_{i=0}^{n-1}$ with $X_i = Y_i \circ O_i$ is a *finite stochastic process* on Ω_T. Associated to this sequence we take the union of their images $\cup_{i=0}^{n-1} \Omega_{X_i} = S \subset \mathbb{R}$ to be the *state space* of the stochastic process and refer to its elements as *states.*

The probability of a value of an element of a finite stochastic process is, in general, somewhat complicated. Indeed, let us consider X_1. Observe that, by definition,

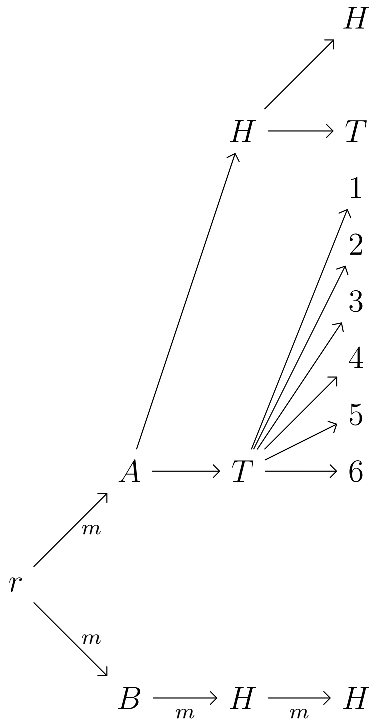
$$P(X_1 = j_1) = \sum_{b \in \Omega_T \ | X_1^{-1}(j_1) \in b} P_B(b)$$

We extend this observation in the straightforward way to $P(X_2 = j_2 \mid X_1 = j_1)$ and, more generally,

$$P(X_n = j_n \mid X_1 = j_1, X_2 = j_2, \ldots, X_{n-1} j_{n-1}) =$$
$$\frac{P(X_1 = j_1 \cap X_2 = j_2 \ldots X_n = j_n)}{P(X_1 = j_1)P(X_2 = j_2 | X_1 = j_1) \cdots P(X_{n-1} = j_{n-1} \, | X_1 = j_1, \ldots, X_{n-2} = j_{n-2})} =$$

In the next section we shall introduce an axiomatic simplification of finite stochastic processes that, in particular,considerably simplifies the above general computation. Stochastic processes together with this axiomatic simplification shall characterize the remainder of the problems we consider in this text.

Now let's do this big example.



6.4 Markov Processes and Markov Chains

In this section we establish the important link between Markov process and linear algebra. The establishment of this link results in the introduction of the Markov chain. To this end, we first consider the stochastic processes distinguished by an additional axiom that simplifies their structure. In applications, this simplification is often described as forgetfulness in the stochastic process. The sense in which the axiom forgets anything is the sense in which it asserts one can know the probability a process is presently in a state only in terms of the previous state, thereby forgetting the prior states which the process had occupied altogether.

This assertion drastically simplifies the complexity inherent in arbitrary stochastic processes and so motivates the following fundamental definition. We say the following definition is the *Markov axiom.*

Definition 51. *Let (Ω_T, β_T, P_B) together with $\{X_i\}_{i=0}^{n-1}$ be a finite stochastic process. Then we say it is a Markov process provided*

$$P_B(X_i = s_i \,|\, X_{i-1} = s_j, X_1 = s_1, X_2 = s_2, \ldots, X_{i-2} = s_{j-1}) = P_B(X_i = s_i \,|\, X_{i-1} = s_j)$$

This axiom means that the probability a stochastic process is in state s_i at time i given its occupancy in the states $s_1, s_2, \ldots, s_j$ at times $1, 2, \ldots, i-1$ can be computed entirely in terms of its conditional probability of being in a prior state, alone. Indeed, as emphasized above, the axiom seems to forget the stochastic process' occupation in the prior states. Intuitively, one would *not* expect such an assertion to be true, in general, in an experiment that occurs over time. However, that is precisely the strength of the defining feature of a Markov process and it is why they are so desirable in applications.

The quantity $P_B(X_i = s_i \,|\, X_{i-1} = s_j)$ in the definition of Markov process is of special significance in what follows, for it determines a matrix representation of T. Define the probability given by the axiom to be $p_{ij} = P_B(X_i = s_i \,|\, X_{i-1} = s_j)$ the *transition probability* from state s_j to state s_i. Then we have the matrix representation $\gamma : \epsilon(E_T) \to \mathbb{R}$ defined by $\gamma(e) = p_{ij}$ for $e = \{v_j, v_i\}$. The corresponding matrix $M = \{p_{ij}\}_{1 \le i \le n, 1 \le j \le n}$ is called the *stochastic matrix* of the Markov process. In particular, we will also consider the matrix representation of the subgraph of all height one edges. We have

$$x_0 = \begin{bmatrix} P_B(X_0 = s_1) \\ P_B(X_0 = s_2) \\ \vdots \\ P_B(X_0 = s_n) \end{bmatrix} = \begin{bmatrix} \delta_{ij} \\ \delta_{ij} \\ \vdots \\ \delta_{ij} \end{bmatrix}$$

where δ_{ij} is the Kroenecker delta, which is 1 if $i = j$ and 0 otherwise. The non-zero value of the Kroenecker delta is determined by the subscript of the state the Markov process begins at, or the state it initially occupies. It is for this reason that x_0 is called an *initial vector*, for it enumerates in its entries the probabilities that the Markov process occupies one of the various states in S at time 0 in the Markov process.

To entirely translate a Markov process into the convenient form of linear algebra, we have at our disposal already a state space S, a stochastic matrix M, and an initial vector x_0. What we must do to finish this translation is to translate the axiom that distinguishes which stochastic processes are Markov processes into linear algebra itself. Together with the following algebraic characterization of the Markov axiom, we shall obtain what we call below a *Markov chain.* To accomplish this aim, we have the following theorem.

Theorem 35. *Let (Ω_T, β_T, P_B) together with $\{X_i\}_{i=0}^{n-1}$ be a Markov process, then the transition probabilities p_{ij} satisfy the relation*

$$P_B(X_i = s_j) = \sum_i P_B(X_{i-1} = s_i) p_{ij}$$

The proof requires the law of total probability. Observe that the fibres of X_{i-1} partition β_T, so that, by the law of total probability, we have

$$\begin{aligned} P_B(X_i = s_j) &= \sum_i P_B(X_{i-1} = s_i \cap X_i = s_j) \\ &= \sum_i P_B(X_j = s_j | X_i = s_j) P_B(X_{i-1} = s_i) \\ &= \sum_i P_B(X_{i-1} = s_i) p_{ij} \end{aligned}$$

where the last equality is obtained by invoking the Markov axiom to simplify the prior line. This argument demonstrates the theorem.

Continuing toward our goal of codifying the Markov axiom in the category of linear algebra, we set

$$x_i = \begin{bmatrix} a_1 = P_B(X_i = s_1) \\ \vdots \\ a_j = P_B(X_i = s_j) \\ \vdots \\ a_n = P_B(X_i = s_n) \end{bmatrix}$$

the column vector whose entries are the probabilities the Markov process in its various states at height i or, hereafter, *time* i. We have enough material now to motivate the following the definition.

Definition 52. *Let (Ω_T, β_T, P_B) together with $\{X_i\}_{i=0}^{n-1}$ be a Markov process. Then we say (S, M, x_0) together with $\{x_i\}_{i=0}^{n-1}$ is a time dependent Markov chain, where S is the state space of the Markov process, M is its stochastic matrix, and x_0 is an initial probability vector.*

To complete our objective of translating the probabilistic concept of a Markov process into the algebraic concept of a Markov chain in a manner that preserves the probabilistic conclusions of the former, we have the following final theorem.

Theorem 36. *Let (Ω_T, β_T, P_B) together with $\{X_i\}_{i=0}^{n-1}$ be a Markov process and (S, M, x_0) together with the sequence $\{x_i\}_{i=0}^{n-1}$ of column vectors of the concomitant Markov chain. Then the Markov axiom*

$$P_B(X_i = s_i \,|\, X_{i-1} = s_j, X_1 = s_1, X_2 = s_2, \ldots, X_{j-2} = s_{j-2}) = P_B(X_i = s_i \,|\, X_{i-1} = s_j)$$

is equivalent to

$$Mx_i = x_{i+1}$$

for $0 \leq i \leq n-1$.

Before proving this, we prematurely state a well-known corollary obtained by applying the theorem recursively.

Corollary 6.1. *With notation as above, $Mx_i = x_{i+1}$ is equivalent to $M^{i+1}x_0 = x_{i+1}$ for $0 \leq i \leq n-1$*

As we write, we mention this corollary prematurely in case the reader is already familiar with this superficially distinct formulation of the Markov axiom in the context of Markov chains. In any case, to prove the theorem, we proceed by induction. Let us establish the base of our induction on i for $i = 0$. Observe then that we must show $Mx_0 = x_1$, where

$$x_1 = \begin{bmatrix} a_1 = P_B(X_1 = s_1) \\ \vdots \\ a_j = P_B(X_1 = s_j) \\ \vdots \\ a_n = P_B(X_1 = s_n) \end{bmatrix}$$

To do this, we shall focus on demonstrating the identity for each entry a_j for $1 \leq j \leq n$. We have

$$a_j = P_B(X_1 = s_j) = \sum_{j=0}^{n-1} P_B(X_1 = s_i \,|\, X_0 = s_j) P_B(X_0 = s_j)$$

by the law of total probability since the fibres of the random variable X_0 induces a partition of β_T. Therefore, by the previous theorem's result, we have

$$a_j = \sum_{j=0}^{n-1} P_B(X_0 = s_j) p_{ij}$$

which is the product of M and x_0 by the definition of the product of a matrix and a vector. Carried out for each j, this argument establishes both $Mx_0 = x_1$ and the basis for induction on i. So, we can assume the result is true for i and demonstrate it follows for $i+1$. We leave the details of this to the reader, which completes the proof. The triple of linear algebraic data attached to a Markov process is the objective of this section, codified in the following definition.

Definition 53. *Let (S, M, x_0) be the triple concomitant to a Markov process and $\{x_i\}_{i=1}^n$ the induced sequence of column vectors satisfying the recurrence relation*

$$Mx_i = x_{i+1}$$

Then we say (S, M, x_0) is a time dependent Markov chain. We say, for

$$M = \{p_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq n}$$

that the transition probability p_{ij} is the probability the chain transitions from state s_j to state s_i. Moreover, for $1 \leq i \leq n$ and

$$x_i = \begin{bmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_n \end{bmatrix}$$

we say a_j is the probability the Markov chain occupies state s_j at time i. Finally, we say (S, M, x_0) is simply a Markov chain if it is time independent. In terms of the induced sequence of probability vectors, we have $\{x_i = M^i x_0\}_{i=0}^{\infty}$ when the Markov chain is time independent.

We shall only consider time independent Markov chains for the remainder of this text. Now do some examples. **EXAMPLES**
Example 6.4.1 Write some stuff *Q.E.F*
Example 6.4.2 Write some stuff *Q.E.F*
Example 6.4.3 Write some stuff *Q.E.F*

6.5 Communication Classes of Markov Chains and Random Walks

In this section, we define the communication classes of a Markov chain and look to the most famous examples of Markov chains as important applications that rely upon the definition in order to justify their introduction. Later in this chapter we shall return to this definition to motivate the page rank algorithm.

Definition 54. *Let (S, M, x_0) be a Markov chain and $s_i, s_j \in S$ be states thereof. We say s_j communicates with s_i if there exists some powers k and l such that both the (i, j) entry in M^k and the (j, i) entry in M^l are strictly positive.*

Given this definition, we can define a relation $R \subset S \times S$ on the state space S of the Markov chain by the rule

$$s_j R s_i \text{ if and only if } s_j \text{ communicates with } s_i$$

We have the following lemma.

Lemma 6.2. *Let (S, M, x_0) be a Markov chain. Then the relation R on S given by $s_j R s_i$ if and only if s_j communicates with s_i is an equivalence relation.*

Let us show that R is reflexive by noting that taking both l and k to be 0 we have that $M^0 = I$ the identity matrix, so that R is reflexive. Next, the relation R is defined to be symmetric by the definition of communication, so R is symmetric. Lastly, it is a routine exercise to show that R is transitive. Suppose s_h communicates with s_j, so that there exist k' and l' satisfying the definition of R. Then one can show that the (i, h) entry of $M^{k'+k}$ and (h, i) entry of $M^{l'+l}$ are strictly positive given that the relvant entries of M^k and M^l are strictly positive, as well. The proof of this lemma invites the following definitions.

Definition 55. *Let (S, M, x_0) be a Markov chain and R the equivalence relation on the state space S given by s_j communicates with s_i. Then we say the equivalence class of s_j with respect to R, say $\mathscr{C}_{s_j}$, is the communication class of s_j. Furthermore, given s_j, we say*

1. *The state s_j is absorbing if $\mathscr{C}_{s_j} = \{s_j\}$*

2. *The state s_j is reflecting if $\mathscr{C}_{s_j} = \{s_{j+1}\}$ or $\{s_{j-1}\}$*

3. *The state s_j is simply transitive if $\mathscr{C}_{s_j} = S$. Equivalently, we say (S, M, x_0) is a regular Markov chain if there exists a simply transitive state.*

According to the results of chapter 4, we can partition S into communication classes. The effect this has is to decompose the attendant stochastic matrix M into block matrix form, where a block corresponds to a communication class. We shall examine this effect below through random walks, which are special cases of Markov chains. We may refine these cases by dividing random walks into absorbing or reflecting walks following the prior general definition.

We define a random walk to be a Markov chain concomitant to a Markov process defined with respect to a tree T with only a single branch. This vague description implies the following definition.

Definition 56. *Let (S, M, x_0) be a Markov chain, such that $|S| = n,$. Then we say it is a random walk if*

1. *S is a totally ordered set, that is $s_1 < s_2 < \cdots < s_{n-1} < s_n$ is the total order on S*

2. *Given $s_i \in S$ for $i = 2, 3, \ldots, n-1$, the transition probabilities are $p_{i,i+1} = q$ and $p_{i+1,i} = p$, where $0 < p < 1$ and $q = 1 - p$*

3. *We say it is an absorbing random walk if the least and greatest states with respect to its total order are absorbing states and we say it is a reflecting walk if the same states are reflecting states instead*

In terms of the stochastic matrix M, we can distinguish absorbing random walks from reflecting random walks according to the transition probabilities in the first and last columns. To wit, if (S, M, x_0) is an absorbing random walk, then its stochastic matrix is given by the $n \times n$ matrix

$$M = \begin{bmatrix} 1 & p & 0 & \cdots & 0 & 1 \\ 0 & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & \ddots & 0 & 0 \\ 0 & 0 & q & 0 & p & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & q & 1 \end{bmatrix}$$

and if (S, M, x_0) is a reflecting random walk, then its stochastic matrix is given by

$$M = \begin{bmatrix} 0 & p & 0 & \cdots & 0 & 1 \\ 1 & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & \ddots & 0 & 0 \\ 0 & 0 & q & 0 & p & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 & 1 \\ 0 & 0 & 0 & 0 & q & 0 \end{bmatrix}$$

Notice that random walks of either class are not regular Markov chains. Indeed, the above stochastic matrices are in block form for distinct communication classes. In terms of the underlying Markov process, we may visualize a random walk as a tree with a single branch, where the root corresponds to the state s_1 and the only leaf corresponds to the state s_n. The branch vertices correspond to the states s_2 through s_{n-1}. We label the transition probability as i increases above the connecting edge and below the connecting edge as i decreases to indicate the matrix representation of the branch for $i = 2, 3, \ldots, n-1$. We can indicate whether the least and greatest states are absorbing by loops and the appropriate transition probability labeling. Observe that

$$\circlearrowleft^1 \; s_1 \underset{0}{\overset{q}{\longleftrightarrow}} s_2 \underset{p}{\overset{q}{\longleftrightarrow}} \cdots \underset{p}{\overset{q}{\longleftrightarrow}} s_{n-1} \underset{0}{\overset{q}{\leftrightarrow}} \circlearrowleft^1 \; s_n$$

illustrates an absorbing walk. Similarly, we can indicate a reflecting random walk by the appropriate below edge labeling. Observe.

$$s_1 \underset{1}{\overset{q}{\longleftrightarrow}} s_2 \underset{p}{\overset{q}{\longleftrightarrow}} \cdots \underset{p}{\overset{q}{\longleftrightarrow}} s_{n-1} \underset{1}{\overset{q}{\longleftrightarrow}} s_n$$

illustrates a reflecting random walk.

Now do some examples. **EXAMPLES**
Example 6.5.1 Write some stuff *Q.E.F*
Example 6.5.2 Write some stuff *Q.E.F*
Example 6.5.3 Write some stuff *Q.E.F*

6.6 Random Walks on Finite Graphs

In this section we define a random walk on a graph to be a Markov chain concomitant to a Markov process defined with respect to a finite graph G. In this sense, the material of this section generalizes that of the previous section if one regards finite graphs as more general than single branched trees. Accordingly, in general, such Markov chains will be irregular, as it is possible for finite graphs to possess both loops and isolated vertices. Such features of a finite graph are translated into Markov chains as a non-trivial partition of the state space into communication classes.

The key application of this material below is to the world wide web and, in particular, to describing a page ranking algorithm for the same by modeling webpages online as vertices and communication classes as hyperlinks between the same. We can rank how popular websites are, in a sense, when the underlying Markov chain is regular. However, as in general there exist obstructions to the regularity of Markov chains predicated upon finite graphs, we shall furnish an algorithm for regularizing such chains. Before we explore this algorithm and its meaning in the final section of this chapter, we have the following definition.

Definition 57. *Let (S, M, x_0) be a Markov chain and $G = (V, E)$ be a finite graph. We say it is a random walk on G if*

1. *The state space $S = V$ is given by the vertices of G*

2. The stochastic matrix is determined by the reciprocal adjacency matrix representation of G*. Specifically, for* $M = [m_1 \ \cdots \ m_j \ \cdots \ m_n]$

where

$$m_j = \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{bmatrix}$$

$a_i = \dfrac{1}{adj(v_j)}$ *if there exists* $e \in E$ *such that* $\epsilon(e) = \{v_j, v_i\}$ *and 0 otherwise, for* $1 \leq i, j \leq n$*.*

Recall from section 6.2 that $adj(v_j) = 1$ if v_j is an isolated vertex in G. We maintain this hypothesis in the above definition of a random walk on a finite graph, G. One notes that the columns of M consist of probability vectors, that is, the sum of their entries is 1, by the definition of the reciprocal adjacency matrix representation. In general, M shall be in block form, according to whether $S = V$ has a non-trivial partition into communication classes. The reciprocal adjacency matrix representation preserves the partition of S. In particular, isolated vertices and cycles in G will obstruct the corresponding random walk on G from being a regular Markov chain. We shall explore the implications of this observation in the following section and learn there of an algorithm to attach a regular Markov chain to an arbitrary finite graph G.

A second random walk related to a finite graph G is obtained when G is a directed graph. One modifies the above definition to $a_i = \dfrac{1}{adj(v_j)}$ if there exists a directed edge $e \in E$ such that $\epsilon(e) = \{v_j, v_i\}$, with $v_j < v_i$, and 0 otherwise, for $1 \leq i, j \leq n$. As an arrow, one insists that v_j is the tail of the arrow in the directed graph G. Naturally, we say such a Markov chain is a random walk on a directed graph G.

Now do some examples. **EXAMPLES**

Example 6.6.1 Write some stuff *Q.E.F*

Example 6.6.2 Write some stuff *Q.E.F*

Example 6.6.3 Write some stuff *Q.E.F*

6.7 Regular Markov Chains

Let (S, M, x_0) together with the sequence of probability vectors $\{x_j = M^j x_0\}_j$ be a Markov chain. We say $\lim_{j\to\infty} x_j = \sigma$ is a steady state vector of the Markov chain provided the limit exists. Equivalently, we say a probability eigenvector associated to the eigenvalue 1 for the stochastic matrix, viz. $\sigma \in \text{Null}(M - I)$ is a steady state vector for (S, M, x_0). One recognizes how these definitions are equivalent according to the naive computation $\lim\limits_{j\to\infty} x_j = \lim\limits_{j\to\infty} M^j x_0 = M \lim\limits_{j\to\infty} x_{j-1} = M\sigma = \sigma$.

We interpret the entries of σ with respect to the Markov chain in two ways. First, the entries of σ are interpreted as the long run probabilities of the Markov chain, which is to say, they reflect the probabilities the Markov chain resides in its various states after an indefinite period. To abuse notation for insight, one could write $x_\infty = \sigma$ to remember this interpretation. Second, as j is arbitrarily large, one could interpret the entries of σ as the proportion of time the Markov chains resides in its various states. It is according to this second interpretation that one obtains the page ranking algorithm.

Theorem 37. *Let (S, M, x_0) be a Markov chain and suppose $M\sigma = \sigma$. Then*

$$\lim_{j\to\infty} M^j = [\sigma \;\; \sigma \;\; \cdots \;\; \sigma]$$

where $[\sigma \;\; \sigma \;\; \cdots \;\; \sigma]$ is the $n \times n$ matrix whose columns are σ.

So the theorem shows that, given the existence of a steady state vector, one is able to compute high powers of M. Accordingly, the existence of such σ is a desirable feature of a Markov chain (S, M, x_0). As such, one would like to know conditions under which a a steady state vector exist for a Markov chain. We have the following definition, which is equivalent to the definition of section 6.3 in terms of communication classes of S.

Definition 58. *Let M be a stochastic matrix. Then we say M is regular if there exists some $j \geq 0$ such that M^j has only positive entries.*

When M is the stochastic matrix of a Markov chain and M is regular, one interprets this feature in terms of the Markov chain to mean that one can transition from any one state to another in exactly j stages if j is the minimum value such that M^j has only positive entries. Equivalently, there exists a simply transitive state in S and that its partition with respect to its communication classes is of length one. The following theorem characterizes when this occurs.

Theorem 38. *Let (S, M, x_0) be a Markov chain such that M is regular. Then it is true that*

1. *There exists a steady state vector σ for the Markov chain*
2. $\lim_{j\to\infty} M^j = [\sigma \;\; \sigma \;\; \cdots \;\; \sigma]$

Our main application of this material is the page ranking algorithm. Let (S, M, x_0) be a arbitrary Markov chain. The page ranking algorithm shall associate it to a regular Markov chain, say (S, M_R, x_0), in a canonical way. The justification for this algorithm is that by executing it one is assured that there exists a steady state vector σ with respect to the regularized Markov chain. The interpretation of the entries of its steady state vector σ as the proportion of time the Markov chain occupies its various states can be used to rank the states according to these occupation times. In particular, if we model the internet by a directed graph as in the previous section, regularizing the associated Markov chain allows us to rank its webpages. This application is both famous and lucrative.

Given a random walk on a finite directed graph $G = (V, E)$, say (S, M, x_0), the two obstructions to its regularity we must consider are loops and isolated vertices. The reason for this, we recall, is that they correspond to elementary columns in M obtained from the reciprocal adjacency matrix representation. It follows from matrix multiplication that no matrix with elementary columns is regular. It is for this reason the following algorithm is focused upon these amending these features of a finite graph.

Proposition 22. ***PAGE RANKING ALGORITHM***

Let (S, M, x_0) be a random walk on a finite directed graph $G = (V, E)$. Then the following algorithm replaces its stochastic matrix with a necessarily regular matrix, M_R.

1. *If $v \in V$ is an terminal vertex, replace the corresponding column under the matrix representation of G by the column vector whose entries are all the reciprocals of the cardinality of V. Specifically, if $|V| = n$, then change the column of v into*

$$\begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

2. *Let $0 \leq p \leq 1$ and K be the $n \times n$ matrix whose entries are all identically $\frac{1}{n}$, that is, $k_{i,j} = \frac{1}{n}$ for all $1 \leq i, j \leq n$. Then compute*

$$M_R = pM^{\star} + (1 - p)K$$

To prove this proposition, we simply must show that the dimension of the eigenspace associated to the eigenvalue 1 is greater than 0. That is, $\dim\mathrm{Null}(M_R - I) \neq 0$ But this is clear because .Therefore matrix M_R obtained from this algorithm is regular by the theorem as a steady state vector exists after we rescale. We say that M_R *regularizes* M. The algorithm motivates the following definition.

Definition 59. *Let (S, M, x_0) be a random walk on a finite directed graph $G = (V, E)$. Then we say the Markov chain (S, M_R, x_0) is the regularization of (S, M, x_0), where $M_R = pM^{\star} + (1 - p)K$.*

Now do some examples. **EXAMPLES**
Example 6.7.1 Write some stuff *Q.E.F*
Example 6.7.2 Write some stuff *Q.E.F*
Example 6.7.3 Write some stuff *Q.E.F*

6.8 Probabilistic Automata and Markov Chains

In this section, we emphasize the relationship between so-called finite automata of computer science and finite Markov chains, as presented above.

6.9 Exercises

7 Markov Decision Processes

7.1 Introduction

7.2 Exercises