Chapter 1

Foundations of Probability

1.1 Introduction

What is probability theory? It is an *axiomatic* theory which *describes and predicts* the outcomes of inexact, *repeated* experiments. Note the emphases in the above definition. The basis of probabilistic analysis is to determine or estimate the probabilities that certain known events occur, and then to use the axioms of probability theory to derive probabilities of other events of interest, and to predict the outcomes of certain experiments.

For example, consider any card game. The inexact experiment is the shuffling of a deck of cards, with the outcome being the order in which the cards appear. An estimate of the underlying probabilities would be that all orderings are equally likely; an event might be a collection of outcomes, such as all the outcomes where the ace of spades is the first card. The underlying events would then be assigned a given probability.

Based on the underlying probability of the events, you may wish to compute the probability that, if you are playing alone against a dealer, you would win a hand of blackjack. Certain orderings of the cards lead to winning hands, and the probability of winning can be computed from the combined information on the orderings.

While card games and other games of chance make fun illustrations for applications of probability, we are interested in using probability for engineering problems. Why do we use probability in such problems? First, we use probability to model phenomena whose outcomes are too hard to model because they involve too many microscopic factors. For instance, the temperature in a room is the result of kinetic energy events released from particle collisions, but modeling those events by representing trajectories of molecules in a room requires very large scale computations. Instead, we can use a probabilistic description of those collisions that forms the basis for thermodynamics.

A different reason we use probability in engineering is to model lack of precision in measurement. No measurement instrument is exact, and all measurements incur some degree of error. We use probability for representing the errors in what we measure versus the actual measured value. For instance, data received over communications channels are subject to unknown distortions whose effect is captured using probability models. A third reason to use probability in engineering problems arises in representing physical phenomena at atomic levels, in modern physics and quantum mechanics. Heisenberg's famous uncertainty principle uses probability to describe fundamental limits in knowing both the position and momentum of atomic particles.

Below is a brief list of examples of how probability models are used in different fields of engineering and science:

- Game Theory: We model outcomes of games of chance, such as cards, rolls of dice, landing of roulette balls, etc. We use those models to derive superior playing strategies that maximize our odds of winning.
- Weather: the evolution of weather fronts over time is subject to many unknown variations, so weather prediction uses probability to estimate likely weather patterns, including predicting hurricane trajectories and strenghts.
- Finance: Probability models are the foundation for mathematical finance, to represent the uncertain evolution of stock prices over time.

- Physics: probability is used to represent possible locations for electrons in orbits, and in statistical mechanics to represent macroscopic effects of numerous molecular motions.
- Molecular Biology: potential DNA mutations of a virus are represented using probability models.
- Science and Engineering Measurements: errors are represented using probability models. Max Born made the observation measured values are within a factor from true values. To quote Max Born, one of the pioneers of quantum mechanics, "The conception of chance enters in the very first steps of scientific activity in virtue of the fact that no observation is absolutely correct."
- Circuits: The true resistance, capacitance and inductance of circuit elements is variable, and these variations are often modeled using probability.
- Optics: The actual number of photons per unit time emitted by a source of given intensity is random, and modeled using probability.
- Transportation: The travel time on roads, the routes selected by traffic, and the wait times at intersections and toll booths are represented using probability models to predict traffic flow.
- Manufacturing: Production times of parts, demand for products, variations in supply chain deliveries are effects that are often modeled by probability.
- Robotics: Problems in determining robot position from sensor data are estimation problems solved using the theory of probability
- Medicine: Problems in diagnosis based on observed patient data are fundamental hypothesis testing problems best addressed using probability.
  item Nuclear Engineering: Failure analysis and diagnosis is based on probabilistic reasoning.
- Astronomy: Detecting and tracking the location of celestial objects using different instruments is based on the theory of probability.
- Data Science: The foundations of data science are the probabilistic theories of estimation and classification.

As the above list indicates, Probability Theory is useful across a wide range of engineering applications.

What do we mean by the probability of an event? This foundational question has been the focus of debate for several centuries, and has several possible answers: One interpretation is the *frequentist* interpretation that the probability of an event means that, if an experiment is repeated an infinite number of times, the probability of the event is the fraction of times that the event occurs in the repeated experiments. This is often used when dealing with simple physical processes, such as rolling dice, shuffling cards, and measurement systems, where experiments can be repeated a large number of times at low cost.

There is a different school of thought: the *subjectivist* interpretation of the probability of an event represents an individual belief that the event will occur, and reflects how much one would be willing to bet that the event will occur. This interpretation is most appropriate when experiments cannot be repeated, such as in economics and social situations. For instance, what is the probability that the New England Patriots will win the Super Bowl this year? That is not an experiment that can be repeated; furthermore, asking that question from different individuals can result in very different estimates of that probability. Similarly, the probability that a nuclear reactor will fail corresponds to events that are hard to repeat, and hence are often nothing but subjective estimates.

How are event probabilities estimated? In the *frequentist* approach, we use statistical observations: We perform an experiment a large number of times N, and count the number of times that the event A is observed, as N_A. The ratio of the two, $\frac{N_A}{N}$ is then estimated as the probability that event A occurs when the experiment is conducted. This estimate varies with the number of times you run the experiment! Ideally, you would like to conduct an infinite number of experiments, but that is impractical, and may not even give you a consistent answer. In this course, we will describe a theoretical foundation for this approach, which shows that as $N \to \infty$, the ratio approaches the underlying correct probability of the event A.

For experiments that are hard to repeat, a different approach at determining probabilties of outcomes is to apply some subjective beliefs based on principles of "equality" or "nonprejudice": If there is no reason to believe that some events are more probable than others, assume they are equally probable. This approach is typical for games of chance, where we assumed ideal balanced coins, dice, roulette wheels, etc. In these experiments, the number of outcomes is typically finite, and the probability of an event is proportionate to the number of outcomes of the experiment that are in the event. Thus, the probability that a roll of two six-sided dice totals 7 is proportional to the number of possible dice outcomes that total 7.

This course is based on the modern axiomatic theory of probability, espoused by mathematicians such as Andrey Kolmogorov: Treat any experiment as generating outcomes in a set (finite or not): the sample space. Events are subsets of the sample space, and probability is any function that assigns a number in [0,1] to an event in a *consistent* way that must satisfy some intuitively appealing properties that will be discussed later. Thus, in a subjectivist interpretation, we cannot assign arbitrary probabilities to different events (the probability that the Patriots will win the Super Bowl versus the probability that a different team will win the Super Bowl.) However, as long as the probabilities are assigned consistent with the axioms we will present, we can use the foundations of probability theory to analyze and predict outcomes in engineering applications in a rigorous manner.

In essence, probability theory provides us with a "calculus" for representing and reasoning about uncertainty that is consistent with basic axiomatic foundations. Probability Theory is an axiomatic theory that models uncertainty in a consistent manner for predictions and decisions. It allows for the computation of probabilities for compound events, chaining of events, derived events as well as conditional inferencing and information processing.

A common question is how probability is related to statistics. The two sciences are close: Probability often deals with predicting the likelihood of future events, while statistics often involves the analysis of the frequency of past events. They use similar axiomatic foundations, and the above distinction is not exclusive. Statistics focuses on the analysis of past data, collected from experiments that involve uncertainty, and is used to understand the results of experiments: validity of outcomes, typicality, cause-effect relationships, and correlations. It builds models based on observations. Probability provides the calculus that allows models built from statistics to be used for predictions and inferencing about future events.

The distinction is best highlighted by an anecdote: A probabilist and a statistician walk to a craps table. The probabilist sees the pair of dice and thinks: "Six-sided dice? Assume each face of the dice is equally likely to land face up. Now compute the chances that each possible number is rolled and bet accordingly." The statistician thinks: "Those dice may look OK, but how do I know that they are not loaded? I'll watch a while, and keep track of how often each number comes up. Then I can decide if my observations are consistent with the assumption of equal-probability faces. Once I'm confident enough that the dice are fair, I'll ask my friend the probabilist to tell me how to bet."

To paraphrase a quote attributed to Persi Diaconis, a Stanford professor, "the problems considered by probability and statistics are inverse to each other. In probability theory we consider some underlying process which has some randomness or uncertainty modeled by random variables, and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations."

1.1.1 A Brief History of Probability

Probability and games of chance arise in anecdotes in every ancient civilization, from Asia, Europe, Central and South Africa. Problems such as weather prediction were critical in estimating agricultural output and governed the pricing of commerce. Observers of astronomical events used astrology for subjective predictions of important events. In metallurgy, early makers of tools used formal rules to reason about mixtures of metals in alloys as well as heating and quenching times to strengthen their tools and reduce impurities. Arab mathematicians used permutations and combinations to list all possible Arabic words with and without vowels, and used early statistics concepts such as frequency analysis for statistical inference. However, none

of these early civilizations developed a consistent calculus for manipulating uncertainty across compound events.

It is fitting that the foundations of modern probability arose from a gambling dispute. These foundations were articulated in a series of articles between mathematicians Blaise Pascal and Pierre de Fermat in 1654. Their discussion concerned a game of chance involving multiple rounds with two players who have equal chances of winning each round. The players contribute equally to a prize pot, and agree in advance that the first player to win a certain number of rounds will collect the entire prize, say the first to win five games. Now suppose that the game is interrupted by external circumstances before either player has achieved victory, so that player 1 has won 3 games and player 2 has won 2. How does one then divide the pot fairly?

Pascal and Fermat articulated some desired properties of the solution: a player who is closer to winning should get a larger part of the pot. But the problem is not merely one of calculation; it also involves deciding what a "fair" division actually is. In their discussions, Pascal and Fermat provided a convincing, self-consistent solution to this problem, and also developed concepts that are still fundamental to probability theory.

To them, it was clear that a player with a 7–5 lead in a game to 10 has the same chance of eventually winning as a player with a 17–15 lead in a game to 20, so Pascal and Fermat therefore thought that interruption in either situation should lead to the same division of the pot.

Fermat now reasoned thus:[1] if one player needs r more rounds to win and the other needs s, the game will surely have been won by someone after $r + s - 1$ additional rounds. Fermat was thus able to compute the odds for each player to win, simply by writing down a table of all possible continuations and counting how many of them would lead to each player winning. Fermat now considered it obviously fair to divide the stakes in proportion to those odds.

Fermat's solution was improved by Pascal in two ways. First, Pascal produced a more elaborate argument why the resulting division should be considered fair. Second, he showed how to calculate the correct division more efficiently than Fermat's tabular method, using a recursive technique.

Shortly after, encouraged by Pascal, Christiann Huygens published the first book of Probability that used their axiomatic framework[2]. Because of the appeal of games of chance, probability theory soon became popular, and the subject developed rapidly during the 18th century. One of the major contributors during this period was Jacob Bernoulli, who studied games with uneven odds, and whose work[3] led to the law of large numbers and to the definition of stochastic convergence, which was the foundation for the frequentist approach to probability. His analysis of games led to the modern concept of Bernoulli random variables and the binomial distribution.

Later in the 18th century, mathematician Abraham de Moivre developed a technique for approximating binomial coefficients[4]. de Moivre's work led to the development of the Central Limit Theorem, and the use of the Gaussian distribution as a fundamental tool in probability and statistics.

Most of the early work in probability theory focused on games of chance. In 1812 Pierre-Simon, Marquis de Laplace, introduced many new ideas in his book, *Théorie Analytique des Probabilités*. Laplace applied probabilistic ideas to many scientific and practical problems, such as the theory of errors, statistical mechanics and actuarial mathematics. In a subsequent article[5], Laplace set out the principles for Bayesian reasoning and inference, and developed the use of characteristic functions and moment generating functions for estimation of moments of random variables. He also connected the principles of least squares estimation to probabilistic inferencing. In his book, Laplace wrote "We see that the theory of probability is at the bottom only common sense reduced to calculation; ... The most important questions in life are, for the most part, really only problems of probability."

[1] Keith Devlin: The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern.
[2] Rekeningh in Spelen van Gluck, translated as "On Reasoning in Games of Chance".
[3] Ars Conjectandi, literally translated as 'art of conjecturing', published after his death in 1713.
[4] "Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Seriem expansi," in "The Doctrine of Chance's" (1718).
[5] Essai philosophique sur les probabilités (1814).

Another important 19th century contributor was Siméon Denis Poisson, who was a mathematical physicist working on various electromagnetic and optics problems. Poisson published a memoir in 1830, where he discusses the ratio of female births and male births in France using the theory of Laplace and binomial distributions based on Bernoulli's work. Poisson proves the weak law of large numbers first. Then, he considers a different limit where the number of births n grows, but the probability of a female birth p diminishes so that pn is constant. He introduced the Poisson distribution as the limit distribution in this problem.

In Russia, Pafnuty Chebyshev is one of the founding fathers of Russian mathematics. His contributions to Probability Theory in the 19th century are extensive. He is best known for the the Chebyshev inequality[6] that bounds the probability that a random variable with known mean and standard deviation has an outcome that is more than a given number of standard deviations away from the mean, and is used to prove the weak law of large numbers in a general setting. He was also an academic mentor of Andrey Markov, another major contributor to the development of Probability Theory.

Andrey Markov is the developer of the theory of Markov Chains. He rigorously proved extensions of the central limit theorem and the law of large numbers to sequences of dependent random variables, a problem that he started working on with Chebychev. His extensive contributions are reflected by the many modern concepts that bear his name, including the Markov inequality, Markov chains, Markov processes, the Gauss-Markov theorem, and Markov random fields.

In the early 20th century, Ronald Fisher developed the foundations of modern statistical analysis, including maximum likelihood detection, analysis of variance, design of experiments, and Fisher information. He applied his principles to botany and genetics and became well-known as a biostatistician. Many modern techniques in data science and statistics carry his name, including the Fisher's linear discriminant and the Behrens-Fisher distribution.

In 1933, Andrey Kolmogorov published his book, *Foundations of the Theory of Probability*, laying the modern axiomatic foundations of probability theory that we teach today. Subsequently, Kolmogorov extended his work to develop the foundation for estimation, smoothing and prediction for stochastic processes, key techniques that are at the heart of modern navigation systems. In statistics, he is best known for the Kolmogorov-Smirnov test for testing whether a collection of independent samples corresponds to a given distribution for a random variable.

1.1.2 Probability at Boston University's College of Engineering

We briefly mention some of the research areas at Boston University that use probability as its foundations.

In the areas of communications and network systems, probability is used in the modeling of traffic, and in performance analysis of networks. It is also used in the physical layer processes of designing signaling strategies, along with coding and decoding. It provides the foundation for information theory and the design of efficient coding strategies for wired and wireless systems.

Probability is extensively used in the analysis of manufacturing systems and networks. Dynamic modeling of demand and production involves probabilistic principles. Key techniques for quality control and product development involve important concepts from design of experiments and statistics.

In aerospace and robotics systems, probability theory provides the foundations for estimation of the system conditions such as location and orientation using noisy sensors, so that effective control can be applied. For intelligent systems and autonomy, probability provides the foundations for machine learning algorithms for robot vision, classification and situation assessment.

In acoustics, we model propagation of waves through random media using the principles of probability, as well as in sonar signal detection and imaging. Probability also provides the foundation for acoustic imaging.

[6] P. L. Chebyshev, Des valeurs moyennes,J.Math.Pures Appl.(2), 1867.

In space physics, we use probability principles as the foundation for imaging the ionosphere with incoherent scatter radar, for tracking space objects using telescopes and radars, and for representing uncertainty in propagation of light through the atmosphere in adaptive optics.

In biomedical systems, we use probability to model the effectiveness of treatments and procedures, and to assess risks. We also use probability to reduce noise in signals and extract meaningful information from signals and images.

In signal and image processing, probability provides the foundation for signal enhancement, denoising, detection and inference, including some recent work on imaging with single-photon cameras. In photonics and nanotechnology, probability provides the foundation for photon propagation and quantum mechanics.

In computer engineering, probability is used for the analysis of algorithms, and for reliability, fault detection and isolation. Probability is also used for the design of novel algorithms for solving hard combinatorial problems that exploit randomness.

Figure 1.1 shows examples of some of the applications listed above.

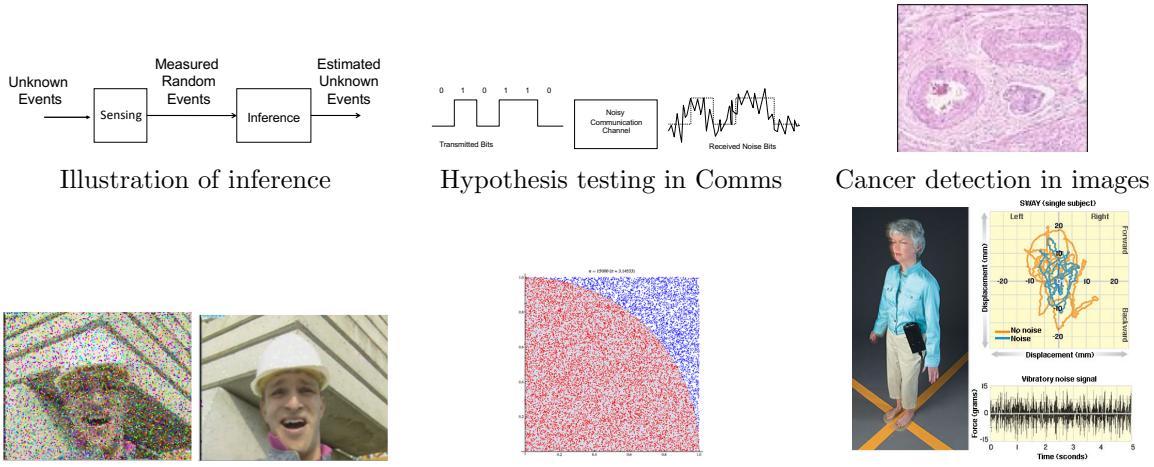


Illustration of inference | Hypothesis testing in Comms | Cancer detection in images

Noise mitigation | Monte Carlo Computation | Noise for stabilization

Figure 1.1: Applications of Probability

1.2 Axioms of Probability

A formal axiomatic theory of probability is necessary to deal with more complex issues such as chaining of events and derived events. At its foundations are fundamental definitions that allow a formulation, along with specific axioms that are accepted without proof as needed by the theory. What follows in the theory of probability are theorems, propositions and lemmas that are consequences of the axioms and definitions and allow the application of probability.

We begin with a review of set theory, which forms the mathematical basis for much of the axioms of probability theory.

1.2.1 Set Theory

Definition 1.1
A set is a collection of elements.

Elements can be anything you like: numbers, letters, people, movies, combinations of items, etc. We usually use capital letters (e.g. A) to denote sets, and lower case letters (e.g. e, x) to denote elements of the set.

A set can be empty, which is called the null set, also denoted by the symbol $\emptyset$. The collection can have a finite number of elements, a countably infinite number of elements, or an uncountable number of elements (e.g. an interval of the real line.) There are several ways to define a set, including

- List its elements: $A = \{1, 3, 5, 7\}$.
- Give a rule in words to generate the set: $A = \{\text{odd integers greater than } 2\}$.
- Give a rule using mathematical symbols: $A = \{x \text{ integer} : x > 2\}$.

In the last version, we use the variable x, and the colon ":" is used as a shortcut for the expression "such that". Hence the last rule reads: A is the set of all numbers such that the number is greater than 2 and the number is an integer. We refer to this version as "set-builder notation."

We use the following notation throughout this text:

- $x \in B$ means that "x is an element of the set B".
- $x \notin B$ means that "x is not an element of the set B".
- The **empty set** or **null set** is the set with no elements. Notation: $\emptyset$ or $\{\ \}$.
- We denote by Ω the **universal set**, i.e. the set of all possible elements.
- A **subset** A of the set B, denoted as $A \subset B$, is a collection of some (or none) of the elements that are in B.
- Two sets are **equal** if $A \subset B$ and $B \subset A$. Thus, the two sets contain the same elements.

A *Venn Diagram* can be used to illustrate relationships between sets. For instance, the figures in 1.2 illustrate Venn diagrams for different set operations. Understanding set operations is much easier if you can visualize the operations using a Venn diagram.

On sets, we define elementary set operations:

- Set complement $A^c = \{x : x \in \Omega \text{ and } x \notin A\}$. Note that $(A^c)^c = A$.
- Set union $A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}$. This is sometimes written as $A + B$.
- Set intersection $A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\}$. This is sometimes written as $A \cdot B$.
- Set Difference $A - B = \{x \in \Omega : x \in A \text{ and } x \notin B\}$. Note $A - B = A \cap B^c$.

These operations are illustrated in 1.2.

Below are other important set concepts that we use in the course:

- A and B are disjoint, or mutually exclusive, sets if and only if $A \cap B = \emptyset$.
- A finite collection of sets $A_1, \ldots, A_n$ are *mutually exclusive* if and only if $A_i \cap A_j = \emptyset$ for any $i \neq j \in \{1, \ldots, n\}$.
- A finite collection of sets $A_1, \ldots, A_n$ is *collectively exhaustive* in Ω if and only if $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$.

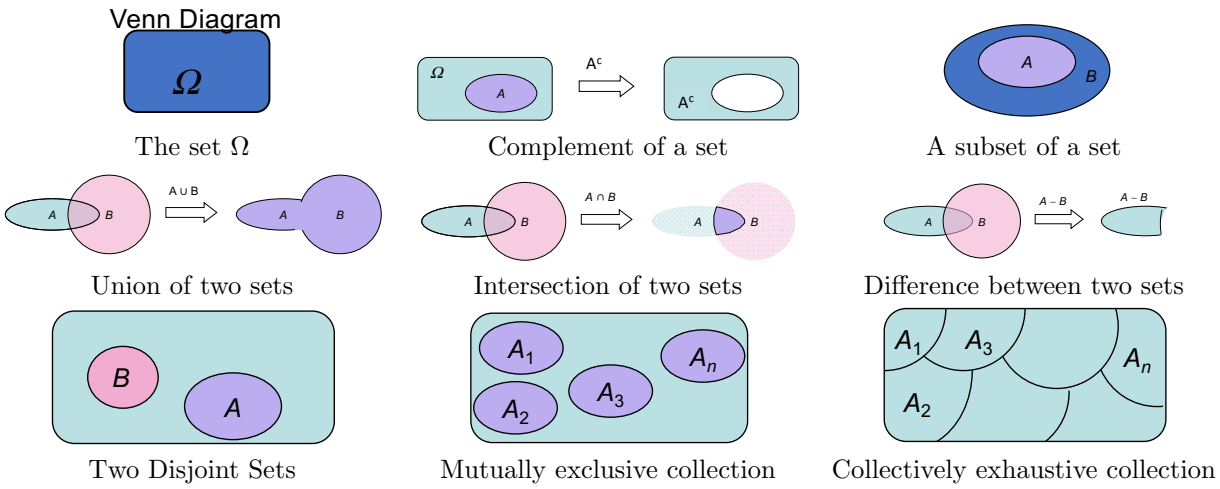


Figure 1.2: Illustration of Set Operations and Concepts

- A countable collection of sets $A_1, A_2, \ldots$, is *mutually exclusive* in Ω if and only if $A_i \cap A_j = \emptyset$ for all $i, j \in \{1, 2, \ldots\}$.

- A countable collection of sets $A_1, A_2, \ldots$, is *collectively exhaustive* in Ω if and only if $A_1 \cup A_2 \cup \ldots = \Omega$.

- A finite or countable collection of sets is a **partition** if it is both mutually exclusive and collectively exhaustive.

From the above definitions, there are several results that can be derived, known as De Morgan's Theorems. The proof of these is obvious; Figure 1.2.1 illustrates the proof of the first result.

- $(A \cup B)^c = A^c \cap B^c$. That is, an element that is not in (A or B) must be (not in A) and (not in b).

- $(A_1 \cup A_2 \cup A_3 \cup \ldots)^c = A_1^c \cap A_2^c \cap A_3^c \cap \ldots$

- $(A_1 \cap A_2 \cap A_3 \cap \ldots)^c = A_1^c \cup A_2^c \cup A_3^c \cup \ldots$

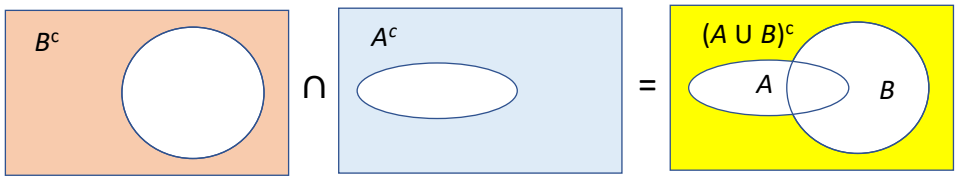


Figure 1.3: Illustration of De Morgan's First Theorem.

To complete this section, we review some mathematical notation that we use throughout these notes. We use the symbol $\forall$ to denote *for all*. Hence $\forall x \in A$ means for all elements x of the set A. The existential qualifier $\exists$ is used to denote that there exists an element. Hence $\exists x \in A$ means that there exists at least one element x that belongs to A. The negative of there exists is denoted $\nexists$.

1.2.2 Probability Axioms

The basic model for probability begins with the concept of a random experiment: An **experiment** is a procedure that generates an observable outcome. An **outcome** is a possible observation of an experiment. The **sample space** Ω of an experiment is the set of all possible outcomes ω of the experiment. Each possible **outcome** ω is an element of the sample space Ω.

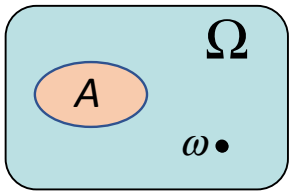


Figure 1.4: Illustration of outcomes ω and events A.

An **event** is subset of Ω: that is, a collection of outcomes. Note that an event may contain a single element, or be the empty set, or be all of Ω. An event is something we will assign probability to; in our axiomatic theory, we assign probabilities to events, not outcomes. Note that it is possible that not every subset of Ω is an event, as we will explain later.

Example 1.1
Experiment: roll a normal six-sided die once. An outcome is the number that shows up on top of the die, which is in $\{1, 2, 3, 4, 5, 6\}$. The sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. Examples of events are $E_1 = \{1, 3, 5\}$, the set of all odd outcomes; E_2 = set of all outcomes greater than 2 ($\{3, 4, 5, 6\}$); and E_3 = set of all outcomes that are the square of an integer ($\{1, 4\}$).

Example 1.2
Experiment: Perform 2 rolls of a quadrilateral (four-sided) die, record both numbers. An outcome is the ordered pair of numbers: $\{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$, so we have Ω consisting of 16 ordered pairs. Examples of events are $E_1\{(1, 3), (2, 2), (3, 1)\}$, the set of all outcomes where the two numbers sum to 4; E_2 = the set of all outcomes that sum to an odd number ($\{(1, 2), (1, 4), (2, 1), (2, 3), (3, 2), (3, 4), (4, 1), (4, 3)\}$.)

Example 1.3
Experiment: Go to the Green Line station on St. Mary's Street and Commonwealth Avenue, going West and wait for the train to arrive. The outcome is the number of minutes (as a real number) before the train arrives. Hence, an outcome is a number x in the sample space $\Omega = [0, \infty)$. Examples of events are $E_1 = \{\text{train arrives under five minutes}\} = \{\omega : \omega < 5\}$; and $E_2 = \{\text{train arrives in more than 20 minutes}\} = \{\omega : \omega > 20\}$.

Example 1.4
Experiment: Measure the arrival time of a pulse, arriving at a random time in the interval $[0, T]$. An outcome is the time of arrival, namely a number $t \in [0, T]$. The sample space $\Omega = [0, T]$ contains an uncountable number of outcomes. Examples of events are $E_1 = \{\omega = T/2\}$ which contains a single outcome, or $E_2 = \{\omega : 0 < a < \omega < b < T\}$ that contains an interval of outcomes.

Example 1.5
As an experiment, pick a point in the unit square $[0, 1] \times [0, 1]$. An outcome ω is the ordered pair consisting of the coordinates of the point, namely a pair $(x, y) \in [0, 1] \times [0, 1]$. The sample space Ω is an uncountable set of ordered pairs $\Omega = \{(x, y) \in [0, 1] \times [0, 1]\}$. Examples of events are $E_1 = \{(1/2, 2/3)\}$ which contains a single outcome, or $E_2 = \{(x, y) \in [0, 1]^2 : x + y \leq 0.2\}$ that contains a region of outcomes.

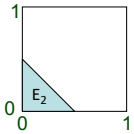


Figure 1.5: Event E_2 in ex. 1.5.

Note that there is a correspondence between the terminology of set theory and that of the probability axioms. We highlight the correspondence in the table below:

Set theory		Probability theory
Universal set	$\Longleftrightarrow$	Sample Space
Element	$\Longleftrightarrow$	Outcome
Subset	$\Longleftrightarrow$	Event

Let's define the collection of all events in an experiment as an event space $\mathcal{E}$. As we highlighted before, we may not want to define every subset of Ω as an event. Since events are sets A for which we want to compute the probability that an outcome is A, there are certain properties that the space of all events must have. We list them below.

Definition 1.2
The event space $\mathcal{E}$ is a collection of subsets of Ω which satisfies the following axioms:

1. $\Omega \in \mathcal{E}$. Thus, the sample space Ω is an event, and the probability that an outcome is in Ω should be 1.
2. If $A \in \mathcal{E}$, then $A^c \in \mathcal{E}$. The complement of an event is also an event, because we want to assign probability to the event that the outcome is not in A.
3. If $A_i \in \mathcal{E}, i = 1, 2, \ldots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{E}$. The union of a countable number of events is an event. This implies that the space $\mathcal{E}$ is closed under the operation of countable unions.

Note that, by the properties of set theory, this implies that $\emptyset \in \mathcal{E}$, because $\Omega^c = \emptyset$. It also implies that the intersection of two events A, B is also an event $A \cap B$, because $A \cap B = (A^c \cup B^c)^c$ using De Morgan's theorems. Also, the set $A - B \equiv \{\omega \in A | \omega \notin B\} \in \mathcal{E}$, because $A - B = A \cap B^c$, which is the intersection of two elements of $\mathcal{E}$.

In terms of assigning probabilities, we will only consider events that are in the event space $\mathcal{E}$. This avoids some mathematical pitfalls that can arise if we consider $\mathcal{E}$ to be all of the subsets of Ω. We call a set *countable* if its elements can be indexed by the natural numbers $0, 1, 2, \ldots$ When the set Ω is countable, we can simply let $\mathcal{E}$ be the collection of all subsets of Ω, as these mathematical difficulties only arise for sets with uncountable numbers of outcomes, such as an interval of real numbers.

The event space $\mathcal{E}$ is often called a **σ-field** (or σ-algebra) in mathematics because it contains Ω, it is closed under countable unions and complementation. In many cases, we construct the set of events $\mathcal{E}$ by specifying some of the basic events that we want to compute probabilities for, and then finding the smallest collection of events that contains the basic events, and is closed under countable unions and complementation.

Example 1.6
Flip 2 coins, a penny and a dime. $\Omega = \{hh, ht, th, tt\}$, with 4 outcomes.

Events of interest $E_i = \{$outcomes with i heads$\}$. Thus, $E_0 = \{tt\}$ contains 1 outcome; $E_1 = \{ht, th\}$ contains 2 outcomes.

What is the smallest event space $\mathcal{E}$ that contains these events? It is $\mathcal{E} = \{\emptyset, E_0, E_1, E_2, E_0 \cup E_1, E_0 \cup E_2, E_1 \cup E_2, \Omega\}$.

Note that $\mathcal{E}$ contains the union of any collections of events, and the complement of each event! However, there are only 8 elements in $\mathcal{E}$, whereas the total number of subsets of Ω is 16. Thus, subsets such as $\{ht\}$ are not events in this event space.

Example 1.7
Consider an experiment consisting of selecting a real number in the interval $[0, 1]$. Consider as events of interest sets of the form $(a, b), a, b \in [0, 1]$. We can define the event space $\mathcal{E}$ as the smallest σ-field that contains these open intervals as events. Note that $\mathcal{E}$ contains the set with two points $\{0, 1\}$ because it is the complement of $(0, 1)$. With further thought, we realize that $\mathcal{E}$ will contain every closed interval $[a, b]$, $a, b \in [0, 1]$, as well as many other events of interest.

An event $A \in \mathcal{E}$ is called an **atom** if it contains only a single outcome; atoms are events of the form $A = \{\omega\}$ for some $\omega \in \Omega$. Events A_i indexed by a set I are called *mutually exclusive* if $A_i \cap A_j = \emptyset$ for all $i, j \in I, i \neq j$. Note that this index set can be infinite in the definition.

We have thus far defined two key components of the axioms of probability: the sample space Ω, which is a collection of outcomes, and the event space $\mathcal{E}$, which is a σ-field collection of subsets of Ω. The third component we need is a *probability measure* $\mathbb{P}$ that assigns a probability value in $[0, 1]$ to each event contained in $\mathcal{E}$; that is, it maps the set of events into the closed unit interval $[0, 1]$. This probability measure $\mathbb{P}[A]$ is interpreted as the probability that the outcome of the experiment is contained in the event $A \in \mathcal{E}$.

The axioms which a probability measure must satisfy are:

1. (**Non-negativity:**) For any event $A \in \mathcal{E}$, $\mathbb{P}[A] \geq 0$ Probabilities are non-negative.
2. (**Normalization:**) $\mathbb{P}[\Omega] = 1$. The probability that we generate an outcome in Ω is one.

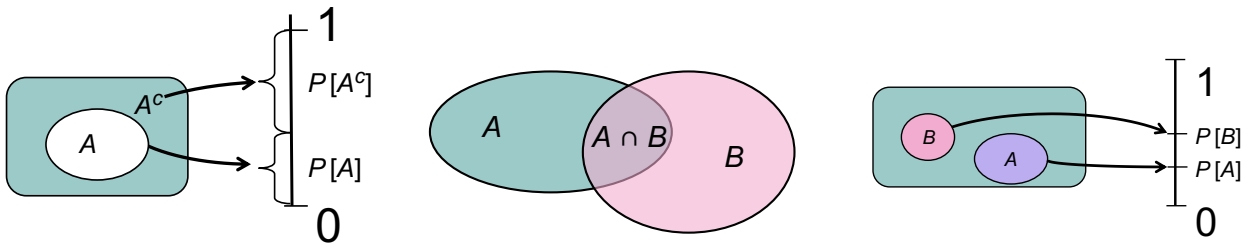
3. (**Countable Additivity**) For any countable collection of mutually exclusive events $A_i, i = 1, 2, \ldots$, we have $\mathbb{P}[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$.

Example 1.8
Consider a sample space $\Omega = [0, 1]$, the unit interval. How do we define events in this space? If we let $\mathcal{E} = \{A : A \subset \Omega\}$, this may become too large a set, with too many elements, such that it is difficult to define a probability measure on these events that satisfy the above axioms.

For instance, assume we want to define a probability measure so that all points have the same probability. For $\omega \in \Omega$, then $\mathbb{P}[\{\omega\}]$ must be a constant. However, that constant must be zero, because otherwise we could find an infinite number of disjoint sets that we could add and get a subset of Ω with infinite probability! Thus, knowing the probability of individual outcomes does not help us in defining a probability measure.

However, consider a different set of events, $E_{a,b} = \{0 \geq a < \omega < b \leq 1\}$. For this interval, we can easily assign a probability measure corresponding to the length of the interval, so that $\mathbb{P}[E_{a,b}] = b - a$. Define the event space $\mathcal{E}$ as the smallest σ-field that contains all the intervals $E_{a,b}$ and is closed under countable unions and complementations: this is known as the *Borel σ-field. Note that every event $A \in \mathcal{E}$ can be written in terms of countable unions and complements of intervals, for which we know how to compute the probability measure. We can extend the measure $\mathbb{P}[A]$ to all elements in $\mathcal{E}$ using the axioms of probability, including the countable additivity axiom. We will show that the countable additivity axiom implies that the probability measure is continuous, and hence we can extend the definition on open intervals to apply to all intervals, and to countable unions and intersections of intervals.*

We are now ready to define a probability space. A *probability space* is a triple $(\Omega, \mathcal{E}, \mathbb{P})$ which is used to describe the outcomes of a random experiment. The set Ω is the set of all possible elementary experiment outcomes ω. The set $\mathcal{E}$ is a $\sigma-$field of events that are subsets of Ω and satisfy the properties of event spaces. The probability measure $\mathbb{P} : \mathcal{E} \to [0, 1]$ satisfies the axioms of probability measures.



$\mathbb{P}[A] + \mathbb{P}[A^c] = 1$ $\quad$ $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$ $\quad$ If Disjoint, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$

Figure 1.6: Illustration of Probability Concepts

Based on the above definition, probability measures can be shown to satisfy additional properties, discussed below. We show brief proofs of selected properties to illustrate how to use the basic properties of probability measures to compute probabilities.

1. $\mathbb{P}[A] = 1 - \mathbb{P}[A^c]$. This follows because A and A^c are mutually exclusive, and $A \cup A^c = \Omega$, so $\mathbb{P}[A] + \mathbb{P}[A^c] = \mathbb{P}[\Omega] = 1$.

2. $\mathbb{P}[\emptyset] = 0$.

3. For any finite collection $A_1, A_2, \ldots, A_n$ of mutually exclusive events,

$$\mathbb{P}[\cup_{i=1}^{n} A_i] = \sum_{i=1}^{n} \mathbb{P}[A_i].$$

4. $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. This follows because $A \cup B = A \cup (B - A)$, and $A, B - A$ are mutually exclusive. Hence, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B - A]$. Furthermore, $B = (B - A) \cup (A \cap B)$, and these two sets are mutually exclusive. Thus, $\mathbb{P}[B] = \mathbb{P}[B - A] + \mathbb{P}[A \cap B]$. Hence, $m[B - A] = \mathbb{P}[B] - \mathbb{P}[A \cap B]$. Substituting into the first equation yields the result.

5. If $B \subset A$, then $\mathbb{P}[B] \leq \mathbb{P}[A]$ and $\mathbb{P}[B] + \mathbb{P}[A - B] = \mathbb{P}[A]$.
6. $\mathbb{P}[A \cup B \cup C) = \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C] - \mathbb{P}[A \cap B] - \mathbb{P}[A \cap C] - \mathbb{P}[B \cap C] + \mathbb{P}[A \cap B \cap C]$;
7. If $A_i, \ldots, A_n$ are mutually exclusive events, $\mathbb{P}[A_i] = \sum_{i=1}^{n} \mathbb{P}[A_i]$.
8. If $A_i, \ldots, A_n$ are mutually exclusive events and $\cup_{i=1}^{n} A_i = \Omega$, then $\sum_{i=1}^{n} \mathbb{P}[A_i] = 1$.
9. If $A_1, A_2, A_3, \ldots$ are mutually exclusive events and $\cup_{i=1}^{\infty} A_i = \Omega$, then for any event A, $\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \cap A_i]$.

This last result is important because it allows us to compute the probability of a complex event as the sum of probabilities of simpler, mutually exclusive events. That is the cornerstone of how most probabilities of events are computed: we break down the events into mutually exclusive subsets for which the probabilities are known, and we use the countable additivity property.

Example 1.9
Consider the example of a shuffle of a deck of cards. The sample space Ω consists of the possible orderings (permutations of 52 cards). While there are many outcomes, there is still a finite number of them. Thus we can make the event space $\mathcal{E}$ the set of all subsets of Ω. Assuming that all permutations are equally likely, the probability measure $\mathbb{P}[A]$ can be defined to be proportional to the number of outcomes in A. For instance, consider the event A consisting of all outcomes where the first card in the deck is the ace of spades. The number of outcomes in A is $51!$ (the first card is the ace of spades, the other 51 can be in any order), and the total number of elements in Ω is $52!$. Hence, $\mathbb{P}[A] = \frac{1}{52}$.

Example 1.10
Consider the toss of a fair coin, with outcomes H, T. The set of outcomes $\Omega = \{H, T\}$. The σ-field $\mathcal{E}$ is

$$\mathcal{F} = \{\{H\}, \{T\}, \emptyset, \{H, T\}\}$$

If the coin is fair, the measure $\mathbb{P}$ will have the following properties:

$$\mathbb{P}[\{H\})] = \frac{1}{2}; \mathbb{P}[\{T\}] = \frac{1}{2}; \mathbb{P}[\{H, T\}] = 1; \mathbb{P}[\emptyset] = 0;$$

One of the important properties of probability measures is the continuity of probability, in the sense specified below. If we have a sequence $A_1 \subset A_2 \subset \cdots$ of increasing events in $\mathcal{E}$, the sequence A_j is monotone increasing and converging to the union $\cup_{i=1}^{\infty} A_i$. Will the probabilities converge also? They will; they are an increasing sequence $\mathbb{P}[A_i]$ of real numbers that are bounded above by one. This allows us to define probability measures on events that can be expressed as limits of events, as shown in the following lemma.

Lemma 1.1
Suppose $A_1, A_2, \ldots$ is a sequence of events in $\mathcal{E}$. Then,

1. If $A_1 \subset A_2 \subset \cdots$, then $\cup_{k=1}^{\infty} A_k \in \mathcal{E}$, $\lim_{k\to\infty} \mathbb{P}[A_k]$ exists, and one defines $\mathbb{P}[\cup_{k=1}^{\infty} A_k] = \lim_{k\to\infty} \mathbb{P}[A_k]$.
2. If $A_1 \supset A_2 \supset \cdots$, then $\cap_{k=1}^{\infty} A_k \in \mathcal{E}, \lim_{k\to\infty} \mathbb{P}[A_k]$ exists, and one defines $\lim_{k\to\infty} \mathbb{P}[A_k] = \mathbb{P}[\cap_{k=1}^{\infty} A_k]$.

proof For the first part, note that $\cup_{k=1}^{\infty} A_k$ is a countable union of events, and hence it is also an event, because event spaces are closed under countable unions and complementation. Let $D_1 = A_1, D_k = A_k - A_{k-1}, k \geq 2$. Note that $D_k \in \mathcal{E}$, because $D_k = A_k \cap A_{k-1}^c$ and intersections of events are also events. Furthermore, the the collection $D_1, D_2, D_3, \cdots$ is mutually exclusive. Then, by the countable additivity axiom of probability,

$$\mathbb{P}[A_k] = \mathbb{P}[\cup_{j=1}^{k} A_j] = \mathbb{P}[\cup_{j=1}^{k} D_j] = \sum_{j=1}^{k} \mathbb{P}[D_j],$$

and thus is an increasing sequence of numbers. Since $\mathbb{P}[A_k]$ is bounded by 1, the monotone convergence theorem guarantees it has a limit that is a number less than or equal to 1, so $\lim_{k\to\infty} \mathbb{P}[A_k]$ exists and is a probability. Thus, the probability of the event $\cup_{k=1}^{\infty} A_k$ is well-defined, as

$$\mathbb{P}[\cup_{i=1}^{\infty} A_k] = \lim_{k\to\infty} \mathbb{P}[A_k] = \sum_{j=1}^{\infty} \mathbb{P}[D_j].$$

For the second part, consider the sets $B_k = A_k^c$. Then, $\mathbb{P}[A_k] = 1 - \mathbb{P}[B_k]$. By the first part, we know $\cup_{i=1}^{\infty} B_k$ is an event, and that $\lim_{k\to\infty} \mathbb{P}[B_k]$ exists and is a probability, and that we define $\mathbb{P}[\cup_{i=1}^{\infty} B_k] = \lim_{k\to\infty} \mathbb{P}[B_k]$.

Now, note $\cap_{k=1}^{\infty} A_k = (\cup_{i=1}^{\infty} B_k)^c$, so it is also an event. Since it is the complement of an event,

$$\mathbb{P}[\cap_{k=1}^{\infty} A_k] = 1 - \mathbb{P}[\cup_{i=1}^{\infty} B_k] = 1 - \lim_{k\to\infty} \mathbb{P}[B_k] = \lim_{k\to\infty} (1 - \mathbb{P}[B_k]) = \lim_{k\to\infty} \mathbb{P}[A_k].$$

Example 1.11
Consider $\Omega = [0, 1]$, the unit interval. Let the set $\mathcal{E}$ be the Borel σ-field in Ω. Note that not all subsets of Ω will be Borel sets, although every interesting subsets we care about is likely to be Borel sets. For an open interval (a, b), we define the measure as its length:

$$\mathbb{P}[(a, b)] = b - a.$$

We can now use the axioms of probability to extend this definition to all Borel sets. It should be easy to see that every Borel set can be written as a countable union of intervals (closed or open, so that a set with only one element x can be written as the interval $[x, x]$). By lemma 1.1, we can now extend the measure $\mathbb{P}[A]$ to compute this uniquely using a limit process.

Why is the concept of event needed over and above the concept of outcome? There are many situations where we want to model the set of possible outcomes as continuous, rather than discrete. In those situations, we know that there are at most a finite number of mutually exclusive events that have probability at least ϵ. By defining probability measures on events, we are able to focus on a finite number of significant events instead of an uncountable number of outcomes. A Furthermore, not every subset of Ω can be considered an event, because it may be impossible to construct a probability measure satisfying the probability axioms. If you are interested in this topic, we show an example in Appendix B of a space with some subsets for which we cannot define a consistent probability measure that satisfy the probability axioms.

In many applications, we define the event space $\mathcal{E}$ by defining a collection of basic events for which we want to compute probabilities, and then finding the smallest σ-field that contains those events. By smallest, we mean the following: A σ-field $\mathcal{E}'$ is said to be a refinement of $\mathcal{E}$ (written as $\mathcal{E} \subset \mathcal{E}'$), if and only if, for any event $A \in \mathcal{E}$, said event is also $A \in \mathcal{E}'$. The smallest or coarsest σ-field that contains a collection of events $\{A_i\}$ is denoted as $\sigma(\{A_i\})$, and is not a refinement of any other σ-field that contains the collection of events $\{A_i\}$. We used this approach to define Borel sets over the unit interval, where the A_i were open intervals in $[0, 1]$. The definition of Borel sets can be generalized to the real line, or n-dimensional Euclidean spaces, or to many other spaces.

As a final note, in any probability space, there can be events which have no probability of occurring. Thus, the difference between two events is often negligible; in such cases, we would like to define a notion of equivalence of events. Two events $A, B \in \mathcal{E}$ are said to be equal with probability one if and only if $\mathbb{P}[A \cup B - A \cap B] = 0$.

The axiomatic theory of probability highlights the approach we need to compute the probability of any event of interest in applications. We outline the steps below, and then proceed to apply to solve probability questions in several examples:

- Identify the sample space from experiment description (the set of all outcomes).
- Describe probability law on events (atoms if finite).
- Identify event of interest
- Calculate the probability of this event as follows:
  - Partition the event of interest into disjoint events for which the probability measures are known.
  - Use axioms of probability to combine the disjoint event probabilities.

Example 1.12
Consider the experiment as one roll of a six-sided die, with balanced outcomes. In this experiment, $\Omega = \{1, \ldots, 6\}$. The problem is to compute the probability of getting an odd outcome (E_1), and the probability of getting an outcome greater than 2 (E_2).

Since we assume the die is balanced, we know $\mathbb{P}[\{\omega\}] = 1/6$, for $\omega \in \Omega$.

We identify $E_1 = \{1, 3, 5\} = \{1\} \cup \{3\} \cup \{5\}$. Given this disjoint decomposition,

$$\mathbb{P}[E_1] = \mathbb{P}[\{1\}] + \mathbb{P}[\{3\}] + \mathbb{P}[\{5\} = \frac{3}{6} = \frac{1}{2}$$

Similarly, $E_2 = \{3, 4, 5, 6\} = \{3\} \cup \{4\} \cup \{5\} \cup \{6\}$ which is another disjoint decomposition, so

$$\mathbb{P}[E_2] = \mathbb{P}[\{3\}] + \mathbb{P}[\{4\}] + \mathbb{P}[\{5\}] + \mathbb{P}[\{6\}] = \frac{2}{3}.$$

Example 1.13
Consider the experiment of 2 rolls of a quadrilateral (four-sided) die, record both numbers and their order. The sample space $\Omega = \{1, 2, 3, 4\}x\{1, 2, 3, 4\}$. We are asked to compute the probability of the following events: E_1 is the set of outcomes that sum to 5. E_2 is the set of outcomes that sum to a prime number not divisible by 3 or 5. E_3 is the set of outcomes such that the smallest of the two numbers is 2.

Proceeding as before, every ordered pair has an equal probability of occurring, which is $1/16$. Now,

$$E_1 = \{(1, 4), (2, 3), (3, 2), (4, 1)\} = \{(1, 4)\} \cup \{(2, 3)\} \cup \{(3, 2)\} \cup \{(4, 1)\}.$$

$$E_2 = \{(1, 1), (3, 4), (4, 3)\} = \{(1, 1)\} \cup \{(3, 4)\} \cup \{(4, 3)\}.$$

$$E_3 = \{(2, 2), (2, 3), (2, 4), (3, 2), (4, 2)\} = \{(2, 2)\} \cup \{(2, 3)\} \cup \{(2, 4)\} \cup \{(3, 2)\} \cup \{(4, 2)\}.$$

Thus,

$$\mathbb{P}[E_1] = \mathbb{P}[\{(1, 4)\}] + \mathbb{P}(\{(2, 3)\}] + \mathbb{P}(\{(3, 2)\}] + \mathbb{P}[\{(4, 1)\}] = \frac{4}{16} = \frac{1}{4}$$

$$\mathbb{P}[E_2) = \mathbb{P}[\{(1, 1)\}] + \mathbb{P}[\{(3, 4)\}] + \mathbb{P}[\{(4, 3))\}] = \frac{3}{16}$$

$$\mathbb{P}[E_3) = \mathbb{P}[\{(2, 2)\}] + \mathbb{P}[\{(2, 3)\}] + \mathbb{P}[\{(2, 4)\}] + \mathbb{P}[\{(3, 2)\}] + m[\{(4, 2)\}] = \frac{5}{16}$$

What about computing $\mathbb{P}[E_1 \cap E_3]$? Note $E_1 \cap E_3 = \{(2, 3), (3, 2)\}$ so $\mathbb{P}[E_1 \cap E_3] = \frac{2}{16}$.

Note that in the above examples, we have used symmetry of the measure $\mathbb{P}$ to simplify computations. The next two examples describe a more complex experiment.

Example 1.14
Our experiment consists of generating telephone calls. Calls can be long or brief, and can voice or data. Thus, an outcome BV denotes a brief voice call, and LD denotes a long data call. The set of outcomes $\Omega = \{LV, LD, BV, BD\}$. The event set $\mathcal{E}$ is the set of all subsets of Ω. Assume we know the following: the probability of a long voice call is 0.35, the probability of a voice call is 0.7, and the probability of a long call is 0.6. What is the probability of a brief data call? What is the probability of a brief voice call? What is the probability of a long data call?

We use the axioms of probability:

$$\mathbb{P}[\{LV, LD\}] = 0.6 = \mathbb{P}[\{LV\}] + \mathbb{P}[\{LD\}] = 0.35 + \mathbb{P}[\{LD\}] \Rightarrow \mathbb{P}[\{LD\}) = 0.25$$

$$\mathbb{P}[\{LV, BV\}] = 0.7 = \mathbb{P}[\{LV\}] + \mathbb{P}[\{BV\}] = 0.35 + \mathbb{P}[\{BV\}] \Rightarrow \mathbb{P}[\{BV\}] = 0.35$$

$$\mathbb{P}[\{BD\}) = 1 - (\mathbb{P}[\{LV\}]] + \mathbb{P}[\{BV\}] + \mathbb{P}[\{LD\}]) = 0.05$$

Example 1.15
Assume we have 2 factories, making the same part. However, they have different quality control. The probability that a part from factory 1 is OK is 0.9. The probability that a part from factory 2 is OK is 0.8.

Assume that factory 1 makes 70% of the parts that are sold, and factory 2 makes 30% of the parts. The outcome from the experiment is a part selected at the store, which will be either good (G) or bad (B), and will come from factory 1 or factory 2.

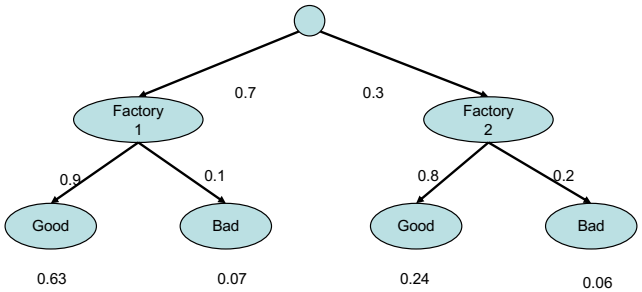


Figure 1.7: Example 1.15.

Given the above description, we can compute the probability of each of the atoms as follows: $\mathbb{P}[\{1G\}] = 0.7{*}0.9 = 0.63$, namely, the probability that the part comes from factory 1 times the probability that the part is good given it comes from factory 1. Similarly, $\mathbb{P}[\{1B\}] = 0.07, \mathbb{P}[\{2G\}] = 0.24, \mathbb{P}[\{2B\}] = 0.06$.

The sample space $\Omega = \{1G, 1B, 2G, 2B\}$. We can view the experiment graphically as shown in Figure 1.15.

What is the probability that the part you buy is good?

$$\mathbb{P}[\{1G, 2G\}] = \mathbb{P}[\{1G\}] + \mathbb{P}[\{2G\}] = 0.87$$

The above example illustrates how our probability calculus allows us to define complex compound experiments.

Note the following: if the sample Ω has a finite number of outcomes $\omega_1, \omega_2, \ldots, \omega_n$, we usually take the event space $\mathcal{E}$ to be the set of all subsets of Ω. In this case, we have the finest disjoint partition of Ω as $\Omega = \{\omega_1\} \cup \{\omega_2\} \cup \cdot \cup \omega_n\}$, and so we can define the probability measure $\mathbb{P}[\cdot]$ on any event by defining it on the atoms $\mathbb{P}[\{\omega_k\}], k = 1, \ldots, n$. In this case, to compute the probability of an event such as $A = \{\omega_1, \omega_3, \omega_5\}$, then we can recognize that A is the union of disjoint sets $\{\omega_1\}, \{\omega_3\}, \{\omega_5\}$, so

$$\mathbb{P}[A] = \mathbb{P}[\{\omega_1\}] + \mathbb{P}[\{\omega_3\}] + \mathbb{P}[\{\omega_5\}].$$

Example 1.16
Consider the following experiment: Ask a person which of the following cities they prefer to live in: Boston, Chicago, Los Angeles, New York, San Francisco. The answer is the outcome, which we denote in shorthand as $S = \{bo, ch, la, ny, sf\}$.

Since it is a finite space, we can assign probabilities to each atom containing a single outcome, as

$$\mathbb{P}[\{bo\}] = \frac{1}{3}; \mathbb{P}[\{sf\}] = \frac{1}{4}; \mathbb{P}[\{ch\}] = \frac{1}{6}; \mathbb{P}[\{la\}] = \frac{1}{8}; \mathbb{P}[\{ny\}] = \frac{1}{8}.$$

Consider the event $A = \{$ east coast city$\}$. Then, $A = \{bo, ny\}$. Using a disjoint decomposition,

$$\mathbb{P}[A] = \mathbb{P}[\{bo\}] + \mathbb{P}[\{ny\}] = \frac{1}{3} + \frac{1}{8} = \frac{11}{24}.$$

Consider the event $B = \{westcoastcity\} = \{sf, la\}$. Then, $\mathbb{P}[B] = \mathbb{P}[\{la\}] + \mathbb{P}[\{sf\}] = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$.

1.3 Conditional Probability and Independence of Events

Conditional probability is the foundation of many engineering applications, especially those involving inference and decision making. Examples of these involve deciding whether radar measurements correspond to the reflections from an aircraft, whether the observation of symptoms indicate the probability that a patient has a disease, and similar questions. In addition, conditional probability is useful for describing complex probability models, such as experiments where the outcome depends on conditions of earlier outcomes. For instance, if you are taking the SATs online, the next question that appears depends on whether you answer the current question correctly or not.

Consider a probability space, and a pair of events $A, B \in \mathcal{E}$ such that $\mathbb{P}[B) > 0$. We define the conditional probability of event A given that B has occurred as:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B)]}{\mathbb{P}[B]}. \tag{1.1}$$

Note that this is not defined when $\mathbb{P}[B)] = 0$. Such events have no probability of being observed in practice, which leads to the lack of a definition. Intuitively, we think of conditioning on event B as **restricting** the universe of possible outcomes to those in B. Hence, only outcomes in $A \cap B$ are now possible out of those in A. Furthermore, we need to **rescale** or normalize so that the conditional probability satisfies the normalization axiom: $\mathbb{P}[B|B] = 1$, which requires that we divide by $\mathbb{P}[B]$.

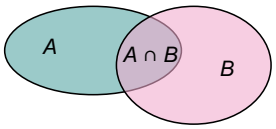


Figure 1.8: Conditional probability.

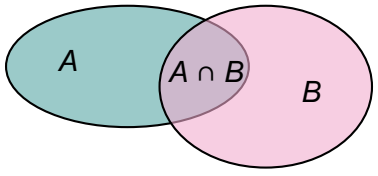
Note the following important relationships:

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A - B] + \mathbb{P}[A \cap B]}$$
$$\mathbb{P}[A \cap B] = \mathbb{P}[B|A]\mathbb{P}[A] = \mathbb{P}[A|B]\mathbb{P}[B]$$

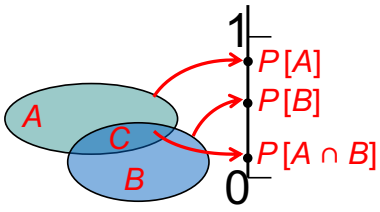
We can extend this to n events $A_1, A_2, \ldots, A_n$ recursively, as:

$$\mathbb{P}[\cap_{k=1}^{n} A_k] = \mathbb{P}[A_1]\mathbb{P}[A_2|A_1]\mathbb{P}[A_3|A_1 \cap A_2] \cdots \mathbb{P}[A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}].$$

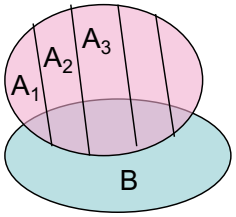
Note that this assumes $\mathbb{P}[A_1 \cap A_2 \cap \cdot \cap A_{n-1}] > 0$ so that conditional probabilities are defined.



$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$ Independence:



$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$



$\mathbb{P}[A|B] = \sum \mathbb{P}[A_i|B]$ for A_i partition A

Figure 1.9: Illustration of Conditional Probability Concepts

Conditional probability functions have an interesting property: they are also probability measures, and a conditional probability space can be defined! Hence, one way of understanding conditional probability is in terms of two operations: Restrict the set of outcomes to B, and compute the relative probability of $A \cap B$ in the restricted sample space B. Restrict means the conditional probability space focuses only on outcomes in B, so the new sample space $\Omega' = B$, and the events are $\mathcal{E}' = \{A \cap B, A \in \mathcal{E}\}$. The conditional probability defines a new measure $\mathbb{P}[\cdot|B]$ on these events that forms a probability space. Rescaling means the original measure $\mathbb{P}[\cdot)]$ must be rescaled (divided by $\mathbb{P}[B]$) so that $\mathbb{P}[\Omega'|B] = \mathbb{P}[B|B] = 1$.

Example 1.17
Consider the city example 1.16. What is the probability that someone's preferred city is Los Angeles, given that their preferred city is on the west coast?

$$\mathbb{P}[\{la\}|\{la,sf\}] = \frac{\mathbb{P}[\{la\} \cap \{la,sf\}]}{\mathbb{P}[\{la,sf\}]} = \frac{\mathbb{P}[\{la\}]}{\mathbb{P}[\{la,sf\}]} = \frac{\frac{1}{8}}{\frac{3}{8}} = \frac{1}{3}.$$

Example 1.18
Consider the previous example 1.17. What is the probability that, if you find the part is good, it was made in factory 1?

The observed event is $B = \{(1G, 2G)\}$, meaning the part is good. The event of interest is $A = \{1G, 1N\}$. We want $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{0.63}{0.87}$.

Here is a brief summary of the properties of conditional probability, which reflect the concept that it is a full probability measure on a restricted sample space corresponding to outcomes in B:

1. $\mathbb{P}[A|B] \in [0,1]$. It is a probability measure.
2. $\mathbb{P}[\Omega|B] = \mathbb{P}[B|B] = 1$.
3. If $A = A_1 \cup A_2 \cup \ldots$ where A_i are mutually exclusive, then
$$\mathbb{P}[A|B] = \mathbb{P}[A_1|B] + \mathbb{P}[A_2|B] + \ldots$$

To show this last item, note the following:

$$\begin{aligned}
\mathbb{P}[A|B] &= \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[(\cup_{i=1}^{\infty} A_i) \cap B]}{\mathbb{P}[B]} \\
&= \frac{\mathbb{P}[\cup_{i=1}^{\infty} [A_i \cap B)]}{\mathbb{P}[B]} \quad \text{(note } A_1 \cap B, A_2 \cap B, \ldots \text{ are mutually exclusive)} \\
&= \frac{\sum_{i=1}^{\infty} \mathbb{P}[A_i \cap B]}{\mathbb{P}[B)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}[A_i \cap B]}{\mathbb{P}[B]} = \sum_{i=1}^{\infty} \mathbb{P}[A_i|B]
\end{aligned}$$

Example 1.19
Consider an experiment where we roll two 6-sided, balanced dice, so all 36 outcomes are equally likely. We consider the following events: B is the set of all outcomes where the smallest of the two numbers rolled is 3. Note that B has 7 elements. A is the set of all outcomes where the first die rolls a 3. Note that A has 6 elements, four of which are in B. In this case,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{4}{7}.$$

Conditional probability is often used to describe the probability measure on complex experiments. In these experiments, you can define the overall probability as a sequence of conditional experiments. We have already seen this illustrated in example 1.15, where the probability that a part was good was dependent on which factory produced it. We illustrate this in the following example:

Example 1.20
We are going to draw three cards from a perfectly shuffled deck of cards, where each order is equally likely. What is the probability that we draw three hearts?

Let A be the event that the first card we draw is a heart. Since there are 13 of those in the 52 cards, $\mathbb{P}[A] = \frac{1}{4}$. Let B be the event that the second card we draw is a heart. Note that we can compute the conditional probability of B given A, because if A was observed, then there are only 12 out of 51 cards left that are hearts, so $\mathbb{P}[B|A] = \frac{12}{51}$. Let C be the event that the third card drawn is a heart. Then, $\mathbb{P}[C|B \cap A] = \frac{11}{50}$.

Then, using the multiplication rule,

$$\mathbb{P}[A \cap B \cap C] = \mathbb{P}[C|B \cap A]\mathbb{P}[B|A]\mathbb{P}[A] = \frac{1}{4} \cdot \frac{12}{51} \cdot \frac{11}{50} = \frac{11}{850}.$$

One of the foundations of inference is Bayes' theorem, which is a consequence of the definition of conditional probability.

Theorem 1.1
Bayes' Rule: Let A, B be events in a probability spae with $\mathbb{P}[A) > 0, \mathbb{P}[B) > 0$. Then,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

Bayes' Rule provides a technique for evaluating the probability of cause, based on the observation of effect. For example, for medical diagnosis, we often to compute $\mathbb{P}[B|A]$, what is the probability of cause B, given the observed effects A. We often have a model for $\mathbb{P}[A|B]$, the probability that certain effects A are associated with cause B. Using Bayes' Rule, we reverse the implication.

An example of the application of Bayes' Rule was seen in example 1.18. In that example, we computed the probability that a good part was manufactured in factory 1, after seeing the effect that the manufactured part was good. However, one of the hardest part for applying Bayes' Rule is computing the denominator $\mathbb{P}[B]$.

One way of doing this is to use the Law of Total Probability, which can be stated as follows.

Theorem 1.2
Let $A_1, A_2, \ldots$ denote a countable set of pairwise mutually exclusive events with $\mathbb{P}[A_i)] > 0$ for $i = 1, 2, \ldots$, and assume that $A_1 \cup A_2 \cup \ldots = \Omega$, so the events are collectively exhaustive. Thus, $A_1, A_2, \ldots$ is a partition. Then, for any event $B \in \mathcal{E}$,

$$\mathbb{P}[B] = \mathbb{P}[B|A_1]\mathbb{P}[A_1] + \mathbb{P}[B|A_2]\mathbb{P}[A_2] + \ldots = \mathbb{P}[B \cap A_1] + \mathbb{P}[B \cap A_2] + \ldots$$

. The Law of Total Probability is useful for computing the denominator $\mathbb{P}[B]$ in Bayes' Rule, by decomposing B as the union of disjoint events.

Example 1.21
You go in for a diagnostic test for a specific diseases, and you test positive! You know that the test can have false positives, and the probability of a false positive (a positive diagnosis when you are not ill) is 0.05. However, you know that the test never misses a disease: the probability that the test returns a positive diagnosis if you are ill is 1.0. However, you know that this is a rare disease, that affects only 0.1% of the population.

Given all that information, what is the probability that you are actually ill?

Let's proceed as before: outcomes of the experiment are $\Omega = \{H+, H-, I+, I-\}$ where H indicates that you are healthy (I for ill), and $+, -$ are the possible outcomes of the diagnostic test. Define the events $H = \{H+, H-\}, I = \{I+, I-\}$ correspond to the person being healthy or ill, and the events $P = \{H+, I+\}, N = \{H-, I-\}$ to the event that the test outcome is positive or negative. Can we construct the probabilities of these outcomes given the information? Note what we are given the following in the problem description:

- Only 0.1% of the population has the disease: $\mathbb{P}[I] = 0.001, \mathbb{P}[H] = 0.999$.
- The probability of a false positive is 0.05: $\mathbb{P}[H \cap P|H] = 0.05, \mathbb{P}[H \cap N|H] = 0.95$.
- The test never has a missed detection: $\mathbb{P}[I \cap P|I] = 1; \mathbb{P}[| \cap N|I] = 0$.

The tree on the right illustrates a conditional diagram for this information, and helps us organize our computation.

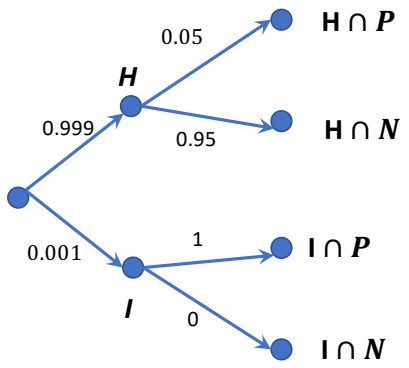


Figure 1.10: Figure for example 1.21.

We want to compute $\mathbb{P}[I|P]$. We use Bayes' rule as:

$$\mathbb{P}[I|P] = \frac{\mathbb{P}[I \cap P]}{\mathbb{P}[P]} = \frac{\mathbb{P}[P|I]\mathbb{P}[I]}{\mathbb{P}[P]}$$

From the diagram, we can see the following: $\mathbb{P}[I] = 0.001, \mathbb{P}[P|I] = 1$, so the terms in the numerator are readily evaluated. What about the denominator? For this, we use the Law of Total Probability, as

$$\mathbb{P}[P] = \mathbb{P}[P|I]\mathbb{P}[I] + \mathbb{P}[P|H]\mathbb{P}[H] = 1 * 0.001 + 0.05(0.999) = 0.05095.$$

Combining the numerator and denominator, we get $\mathbb{P}[I|P] = \frac{0.001}{0.05095} \approx 0.0196$.

Note how we used the Law of Total Probability to compute the denominator, since H, I form a partition of Ω. The message is that you should not be fast to assume you are sick...probability can help understand how to combine the different pieces of information!

Example 1.22
Consider a noisy communication channel, where binary bits are transmitted (values 0 or 1) but received occasionally with errors. Assume that a bit is received correctly with probability 0.95, and is received in error with probability 0.05. Assume that the probability of transmitting a 1 is 0.1, and a zero is 0.9. Given that the bit you received is 0, what is the probability that the transmitted bit was 0?

Define the following events:

Event A_1: Bit 0 transmitted. We are given $\mathbb{P}[A_1] = 0.9$.

Event A_2: Bit 1 transmitted, $\mathbb{P}[A_2] = 0.1$.

Event B_1: Bit 0 received. $\mathbb{P}[B_1|A_1] = 0.95, \mathbb{P}[B_1|A_2] = 0.05$.

Event B_2: Bit 1 received. $\mathbb{P}[B_2|A_1] = 0.05, \mathbb{P}[B_2|A_2] = 0.95$.

We wish to compute $\mathbb{P}[A_1|B_1]$. Using Bayes' Rule, and the Law of Total probability, this is

$$\mathbb{P}[A_1|B_1] = \frac{\mathbb{P}[B_1|A_1]\mathbb{P}[A_1]]}{\mathbb{P}[B_1]} = \frac{\mathbb{P}[B_1|A_1]\mathbb{P}[A_1]}{\mathbb{P}[B_1|A_1]\mathbb{P}[A_1] + \mathbb{P}[B_1|A_2]\mathbb{P}[A_2]} = \frac{0.9 * 0.95}{0.9 * 0.95 + 0.1 * 0.05} \approx 0.994$$

Example 1.23
Monty Hall Game Show: Here is a paradoxical example from the game show "Let's Make a Deal." You know there is a prize behind one of three doors. You are asked to pick one, which you do: door 1. The game show host, Monty Hall, opens door number 2 and shows that there is no prize behind that door. He then gives you a choice to keep your original door, or switch to door 3. Should you switch?

At first glance, the choice seems harmless: There are two doors left, and the prize is behind one of them. It seems like switching and not switching should give you equal chance of winning. However, during the few seasons of the show, those that switched wound up winning 2/3 of the time? Why?

Here is a quick explanation: The original door choice had only 1/3 chance of winning. Hence, switching to both of the other two door choices has 2/3 chance of winning. The fact that Monty opened one of those doors and showed it had no prize means that choosing the other door has the same chance of winning as choosing both doors, namely 2/3.

Let's analyze this using Bayes' Rule: Sample space $\Omega = \{1, 2, 3\}$ corresponding to which door has a prize. The measure on atoms is $\mathbb{P}[\{1\}] = \mathbb{P}[\{2\}] = \mathbb{P}[\{3\}] = 1/3$. Let event $E_1 =$ prize is behind door 1. Let event O be the event that Monty opens a door that does not have a prize behind it. What is $\mathbb{P}[E_1|O]$?

Bayes' Rule: $\mathbb{P}[E_1|O] = \frac{\mathbb{P}[O|E_1]\mathbb{P}[E_1]}{\mathbb{P}[O]}$. Here is the source of the paradox: Monty will always open a door that does not have a prize (or else the game ends!) Hence, $\mathbb{P}[O] = 1, \mathbb{P}[O|E_1] = 1$. Thus, Bayes' Rule yields

$$\mathbb{P}[E_1|O] = \frac{\mathbb{P}[O|E_1]\mathbb{P}[E_1]}{\mathbb{P}[O]} = \mathbb{P}[E_1] = 1/3.$$

which means that not switching wins the prize only 1/3 of the time. Thus, one should always switch!

Example 1.24
3 factories manufacture batteries for an electric car. However, the batteries from factory one meet the needed specification only 70% of the time, while the batteries from factories 2 and 3 meet specifications 80% and 85% of the time respectively. The car manufacturer buys 40% of its batteries from factory 1, 30% of its batteries from factory 2, and the remaining 30% of its batteries from factory 3.

An outcome of this experiment is the battery that is in the car you purchased. The battery was made by one of the three manufacturers, and it either met the specification (denote by G), or did not (denote by B). Thus, the sample space can be described with 6 outcomes, as $\Omega = \{1G, 1B, 2G, 2B, 3G, 3B\}$. Let A denote the event that the battery in your car meets specification. Let B_i denote the event that the battery in your car came from factory i, for $i = 1, 2, 3$. Note that B_1, B_2, B_3 are mutually disjoint and collectively exhaustive. Then, using the Law of Total Probability,

$$\mathbb{P}[A] = \sum_{i=1}^{3} \mathbb{P}[A|B_i]\mathbb{P}[B_i] = 0.7 * 0.4 + 0.8 * 0.3 + 0.85 * 0.3 = 0.775$$

Example 1.25
This is a longer example to show the use of Bayes' Rule and conditional probability. There is a new virus infecting smartphones and randomly compromising some of them. We know that 80% of the smartphones run Android OS, and 20% run a different operating system. Let A denote the event of phones that run Android, and let A^c denote the event of phones that run a different operating system. Let B be the event that the virus infects the phone. We are given that $\mathbb{P}[B|A] = 0.5$, so that $\mathbb{P}[B^c|A] = 0.5$ also. We are also given that $\mathbb{P}[B|A^c] = 1/3$, because non-Android phones are less common.

Let C be the event that the phone is compromised if it is infected. We are given that $\mathbb{P}[C|A \cap B] = 1/4$. The phone can still be compromised even if it is not infected! The probability of this is $\mathbb{P}[C|A\ B^c] = 1/8$. For non-android phones, the probability that the phone is compromised if it is infected is $\mathbb{P}[C|A^c \cap B] = 1/6$, and the probability that the phone is compromised if the phone is not infected is $\mathbb{P}[C|A^c \cap B^c] = 1/8$.

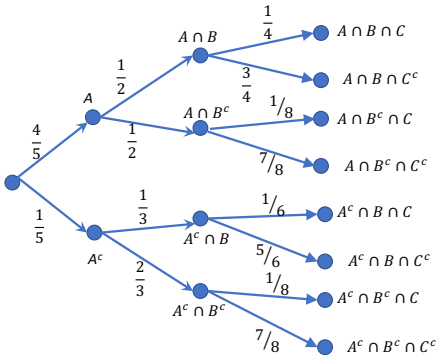


Figure 1.11: Figure for example 1.25.

We can combine all of these probabilities (and their complements) to obtain a complete event diagram that chains these events appropriately. This diagram is shown in Figure 1.11. For instanced, what is the probability that a phone that runs Android has no bug but is still compromised? This is $\mathbb{P}[A \cap B^c \cap C] = \mathbb{P}[A]\mathbb{P}[B^c|A]\mathbb{P}[C|A \cap B^C] = \frac{4}{5} \cdot \frac{1}{2} \cdot \frac{1}{8} = \frac{1}{20}$. Note that this is the product of the probabilities on the branches leading to the end node $A \cap B^c \cap C$.

What is the probability that a phone has a virus but is not compromised? That is $\mathbb{P}[B \cap C^c]$. We can compute this using the Law of Total Probability, as A and A^c form a partition of the sample space. Thus,

$$\mathbb{P}[B \cap C^c] = \mathbb{P}[B \cap C^c|A]\mathbb{P}[A] + \mathbb{P}[B \cap C^c|A^c]\mathbb{P}[A^c] = \mathbb{P}[A \cap B \cap C^c] + \mathbb{P}[A^c \cap B \cap C^c] = \frac{3}{10} + \frac{1}{18} = \frac{16}{45}.$$

Given that a phone is compromised, what is the probability that it is an Android phone? We answer this using Bayes' Rule:

$$\mathbb{P}[A|C] = \frac{\mathbb{P}[C|A]\mathbb{P}[A]}{\mathbb{P}[C]}$$

Note that B and B^c are also a partition of the sample space. Using the Law of Total Probability with the conditional probability $\mathbb{P}[\cdot|A]$, we get

$$\mathbb{P}[C|A] = \mathbb{P}[C|A \cap B]\mathbb{P}[B|A] + \mathbb{P}[C|A \cap B^c]\mathbb{P}[B^c|A] = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{8} \cdot \frac{1}{2} = \frac{3}{16}.$$

Similarly, note that $A \cap B, A \cap B^c, A^c \cap B, A^c \cap B^c$ form a partition of the sample space Y. Then,

$$\begin{aligned}\mathbb{P}[C] &= \mathbb{P}[C \cap A \cap B] + \mathbb{P}[C \cap A \cap B^c] + \mathbb{P}[C \cap A^c \cap B] + \mathbb{P}[C \cap A^c \cap B^c] \\ &= \frac{4}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} + \frac{4}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{5} \cdot \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{5} \cdot \frac{2}{3} \cdot \frac{1}{8} = \frac{8}{45}\end{aligned}$$

1.3.1 Independence

Independence is an important concept in probability. It is one of the most common assumptions used in modeling experiments with multiple sources of randomness, and allows for efficient characterization of the resulting probability measures.

Mathematically, two events A, B in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ are **independent** if and only if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Independence implies $\mathbb{P}[B|A] = \mathbb{P}[B], \mathbb{P}[A|B] = \mathbb{P}[A]$ as long as $\mathbb{P}[A] > 0, \mathbb{P}[B] > 0$, because for any two events, $\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B]$.

Independence has nothing to do with A, B being mutually exclusive, i. e. having no outcomes in common. If A, B have no outcomes in common, then $\mathbb{P}[A \cap B] = 0$, and $\mathbb{P}[A|B] = 0$ also, so it is impossible for $\mathbb{P}[A|B] = \mathbb{P}[A]$ as long as $\mathbb{P}[A] > 0$. If A, B are disjoint, knowing that the experiment outcome is in B implies it is cannot be in A and so this provides information about A, and thus the events are not independent. Independence is a property of the probability measure $\mathbb{P}[\cdot]$ and the specific sets A, B.

If A, B are independent events, then A^c, B are also independent, because

$$\mathbb{P}[A^c \cap B] + \mathbb{P}[A \cap B] = \mathbb{P}[B] = \mathbb{P}[A^c \cap B] + \mathbb{P}[A]\mathbb{P}[B]$$

This implies

$$\mathbb{P}[A^c \cap B] = (1 - \mathbb{P}[A])\mathbb{P}[B] = \mathbb{P}[A^c]\mathbb{P}[B]$$

Example 1.26
Consider a simple sample space $\Omega = \{0, 1, 2, 3\}$ with four elements, and define the probability measure on the atoms to be:

$$\mathbb{P}[\{0\}] = \frac{2}{9}; \mathbb{P}[\{1\}] = \frac{1}{9}; \mathbb{P}[\{2\}] = \frac{4}{9}; \mathbb{P}[\{3\}] = \frac{2}{9}.$$

Define events $A = \{0, 1\}, B = \{2, 3\}, C = \{1, 3\}$.

Are A, B independent? They are mutually exclusive, so they are not, because $\mathbb{P}[A \cap B] = 0$.

Are A, C independent? We have to check: $\mathbb{P}[A] = \mathbb{P}[\{0\}] + \mathbb{P}[\{1\}] = \frac{1}{3}; \mathbb{P}[C] = \mathbb{P}[\{1\}] + \mathbb{P}[\{3\}] = \frac{1}{3}$. $\mathbb{P}[A \cap C] = \mathbb{P}[\{1\}] = 1/9 = \mathbb{P}[A]\mathbb{P}[C] = \frac{1}{3}$. Thus, they are independent!

The concept of independence can be extended to a finite sequence of sets $A_1, \ldots, A_m$, which are mutually independent if

- Any collection of k of the sets $(k < m)$ $A_{j_1}, A_{j_2}, \ldots, A_{j_k}$ are mutually independent. item $\mathbb{P}[A_1 \cap A_2 \cap \ldots \cap A_m) = \mathbb{P}[A_1]\mathbb{P}[A_2] \cdots \mathbb{P}[A_m]$.

.

Note that the above concept of mutual independence implies much more than pairwise independence. For pairwise independence, any two sets A_i, A_j, $i \neq j, i, j \in \{1, \ldots, m\}$ are independent. It is easy to construct examples of events which are pairwise independent, but not mutually independent. Mutually independent means that no subset of the events can be used to predict the probability of occurrence of any of the other events.

Independence can be tedious to check. Often, as in the above example, it is easier to recognize lack of independence, as when two events are mutually exclusive. In many engineering applications, we will assume independence.

Example 1.27
The experiment in this example is to flip a coin twice, and record both faces. The sample space is thus $\Omega = \{HH, HT, TH, TT\}$. Assume the coins are fair, so each atom in the sample space has probability $\frac{1}{4}$. Define the following events: $A =$ {First flip is H}; $B =$ {Second flip is H}; $C =$ {Flips have different outcomes}. Note that $\mathbb{P}[A] = \mathbb{P}[B] = \mathbb{P}[C] = \frac{1}{4}$, as each of the events has two outcomes.

Observe the following: $\mathbb{P}[A \cap B] = \mathbb{P}[\{HH\}] = \frac{1}{4}; \mathbb{P}[A \cap C] = \mathbb{P}[\{HT\}] = \frac{1}{4}; \mathbb{P}[B \cap C] = \mathbb{P}[\{TH\}] = \frac{1}{4}$. Therefore, A and B are independent, and B and C are independent, and A and C are independent! Thus, A, B, C are pairwise independent. However, are they mutually independent?

Note that $A \cap B \cap C = \emptyset$, so $\mathbb{P}[A \cap B \cap C] = 0 \neq \mathbb{P}[A]\mathbb{P}[B]\mathbb{P}[C]$. Therefore, the events A, B, C are not mutually independent.

We now define the concept of conditional independence. Two events A, B are *conditionally independent* given event C if $\mathbb{P}[A \cap B|C] = \mathbb{P}[A|C)]\mathbb{P}[B|C]$. Basically, this defines independence in terms of the conditional probability measure $\mathbb{P}[\cdot|C]$. Note the following: two events that were independent in the original probability measure $\mathbb{P}[\cdot]$ can become conditionally dependent when a third event C is observed as true. Similarly, two events that were not independent originally can become independent given C is observed.

Example 1.28
We illustrate this concept with a simple encryption example. Assume we want to send a single bit M equally likely to be 0 or 1. We generate independently another bit K, equally likely to be 0 or 1, and we send the message

$$C = M \oplus K$$

where $\oplus$ is addition modulo 2. Thus, if $K = 1$, the original bit M is flipped. The sample space of these experiments is $\Omega = \{00, 01, 10, 11\}$ corresponding to possible pairs MK, and $\mathbb{P}[\{ij\}] = 1/4, i, j = 0, 1$.

Define following events are independent: $K_i = \{K = i\}, M_j = \{M = j\}$. Note that K_0, M_1 are independent. Consider the event $C_i = \{C = i\}$.

By construction, K_0 and M_1 are independent, which can be verified by

$$\mathbb{P}[K_0 \cap M_1] = \mathbb{P}[\{00, 10\} \cap \{10, 11\}] = \mathbb{P}[\{10\}] = 0.25 = \mathbb{P}[K_0]\mathbb{P}[M_1]$$

Note that $\mathbb{P}[C_0] = 0.5$. Then,

$$\begin{aligned}
\mathbb{P}[K_0|C_0] &= \frac{\mathbb{P}[K_0 \cap C_0]}{\mathbb{P}[C_0]} = \frac{\mathbb{P}[\{00, 10\} \cap \{00, 11\}]}{\mathbb{P}[C_0]} = 0.5, \\
\mathbb{P}[M_1|C_0] &= \frac{\mathbb{P}[M_1 \cap C_0]}{\mathbb{P}[C_0]} = \frac{\mathbb{P}[\{10, 11\} \cap \{00, 11\}]}{\mathbb{P}[C_0]} = 0.5, \\
\mathbb{P}[K_0 \cap M_1|C_0] &= \frac{\mathbb{P}[K_0 \cap M_1 \cap C_0]}{\mathbb{P}[C_0]} = \frac{m(\emptyset)}{\mathbb{P}[C_0]} = 0.
\end{aligned}$$

which shows that K_0 and M_1 are not conditionally independent given C_0.

Here is another surprising fact: C_0 and M_1 are independent! We show this by computation:

$$\mathbb{P}[C_0 \cap M_1] = \mathbb{P}[\{00, 11\} \cap \{10, 11\}] = 0.25 = \mathbb{P}[C_0]\mathbb{P}[M_1]$$

You can verify that C_i is independent of K_j, and M_k for any set of choices $i, j, k \in \{0, 1\}$! What this implies is that knowing C alone does not reveal anything about M, hence ensuring the privacy of M. In our simple scale, the possible weights are $\{0, 1, 2, 3\}$. In addition to its bias, the scale has an error measuring a weight which is independent for each time you weigh the object, and the error has equal probability in $\{0, 1\}$.

Example 1.29
Here is an example where two dependent events become conditionally independent. An acoustic microphone is listening to detect whether a particular sound waveform is present or not. However, the background noise in the room can either be "loud" or "soft". The probability that the background is "loud" is 0.5. If the background nose is "loud", the microphone will detect the presence of a sound with probability of error 0.4. That is, if the sound is present, the microphone will detect it with probability 0.6, and fail to detect it with probability 0.4.

If the background noise is "soft", the microphone will detect the presence of a sound with probability of error 0.2. The experiment consists of a room with the noise present with background chosen randomly from "loud" or "soft". The microphone will try twice to detect the presence of the sound twice, where the errors the microphone makes are independent for each try., but the background noise is the same in both measurements.

The sample space in this experiment can be stated in terms of 3 variables: l or s for whether the background is loud or soft, $d_1 = 0, 1$ as to whether the first mesurement is a detection, and $d_2 = 0, 1$ as to whether the second measurement is a detection. Thus, the sample space consists of 16 outcomes,

$$\Omega = \{l00, l01, l10, l11, s00, s01, s10, s11\}$$

Given the description of the experiment, we compute the probability of each atom as follows:

$$\mathbb{P}[\{l00\}] = 0.08; \mathbb{P}[\{l01\}] = \mathbb{P}[\{pl10\}] = 0.12; \mathbb{P}[\{l11\}] = 0.18$$

$$\mathbb{P}[\{s00\}] = 0.02; \mathbb{P}[\{s01\}] = \mathbb{P}[\{s10\}] = 0.08; \mathbb{P}[\{s11\}] = 0.32$$

Note that these add up to 1. Define the event A be the event that the first measurement is a detection, and B the event that the second measurement is a detection. By combining the atoms that form these events, we compute $\mathbb{P}[A] = 0.7, \mathbb{P}[B] = 0.7$. However, $\mathbb{P}[A \cap B] = 0.5$, so the events are not independent.

Define the event C to be the event that the background is "loud". By construction, $\mathbb{P}[B) = 0.5$. Now,

$$\mathbb{P}[A|C] = \frac{\mathbb{P}[A \cap C]}{\mathbb{P}[C]} = \frac{0.3}{0.5} = 0.6$$

$$\mathbb{P}[B|C] = \frac{\mathbb{P}[B \cap C]}{\mathbb{P}[C)} = \frac{0.3}{0.5} = 0.6$$

$$\mathbb{P}[A \cap B|C] = \frac{\mathbb{P}[A \cap B \cap C]}{\mathbb{P}[C]} = \frac{0.18}{0.5} = 0.36$$

which shows that A, B are conditionally independent given C, knowledge of the background state. Basically, the dependence in events A, B arises because they contain the common uncertainty from the same background state. This is removed if the background is observed so it is no longer uncertain.

Example 1.30
Consider the experiment of selecting an integer from 1 to 4, where each number is equally likely. Consider the events $\{1, 2\}, \{1, 3\}, \{1, 4\}$; Note that the above events are pairwise independent. However,

$$\mathbb{P}[\{1, 2\} \cap \{1, 3\} \cap \{1, 4\}] = 1/4 \neq \mathbb{P}[\{1, 2\}].\mathbb{P}[\{1, 3\}]\mathbb{P}[\{1, 4\}] = 1/8.$$

Example 1.31
Mutual independence is a much stronger condition that pairwise independence. For example suppose we were looking for 5 genetic markers in blood samples, denoted by A, B, C, D, E. We are given that the presence of marker A is in 1 out of every 100 persons, marker B in 1 of 50 persons, marker C in one of 40 persons, marker D in one of 5 persons and marker E in one of 170 persons. If the presence of each of the markers was mutually independent, the probability that all the markers were present in a blood sample is

$$\mathbb{P}[A \cap B \cap C \cap D \cap E] = \frac{1}{100 * 50 * 40 * 5 * 170} = \frac{1}{170,000,000}$$

However, if all we knew was that the presence of the markers was pairwise independent but not mutually independent, then all we can say is

$$\mathbb{P}[A \cap B \cap C \cap D \cap E) \leq \mathbb{P}[A \cap E) \leq \frac{1}{17,000}$$

Those three orders of magnitude matter!

The concepts of independence and conditional independence are used extensively in this course to construct complex compound experiments. We have already seen this in several examples previously. These experiments are sequences of sub-experiments, where the later subexperiments depend on the outcomes of the earlier subexperiments. That is, we are given an initial subexperiment with events of outcomes A_i, defined by probability measure $\mathbb{P}[A_i]$. Then, we define the next sub-experiment with events of outcomes B_j, defined by conditional probability measures $\mathbb{P}[B_j|A_i]$ depending on the events observed in the earlier subexperiments. Note that we can now define a probability measure on the compound experiment with

$$\mathbb{P}[B_j \cap A_i] = \mathbb{P}[B_j|A_i]\mathbb{P}[A_i]$$

This is particularly simple when the experiments have discrete outcomes and we define the probability measures on atoms. We illustrate this below with two examples.

Example 1.32
We have three factories B_i making a product. The experiment will generate a sample product, which may or may not be acceptable. As a first step, we select which factory will make the product. We describe this by a probability measure over the atoms of this first step, $\mathbb{P}[\{B_i\})$, with probability 0.3, 0.4, 0.3 for each of the outcomes B_1, B_2, B_3. In the next part of the experiment, the selected factory makes a sample product, which may turn out to be acceptable A or not N. We describe this with conditional probability based on the factory selected:

$$\mathbb{P}[A|B_1] = 0.8, \mathbb{P}[N|B_1] = 0.2; \mathbb{P}[A|B_2] = 0.9, \mathbb{P}[N|B_2] = 0.1; \mathbb{P}[A|B_3] = 0.6, \mathbb{P}[N|B_3] = 0.4.$$

We can illustrate this example with a tree diagram, as illustrated in Figure 1.12. The compound experiment has defined all the probabilities.

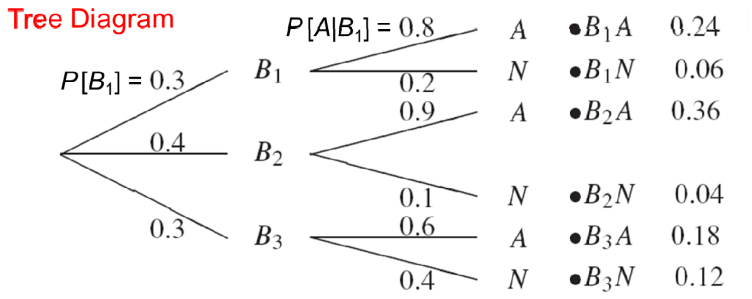


Figure 1.12: Tree diagram for example 1.32.

Example 1.33
Consider a communication channel where we are going to send a four-bit sequence of bits $a_1a_2a_3a_4$. Each bit is generated independently from $\{0, 1\}$ with $\mathbb{P}[\{a_i = 0\}] = 0.4$. The bits are input into the communication channel one at a time, where each bit can be flipped independently by the channel with probability 0.2, or left as is with probability 0.8.

This experiment is illustrated in figure 1.13.

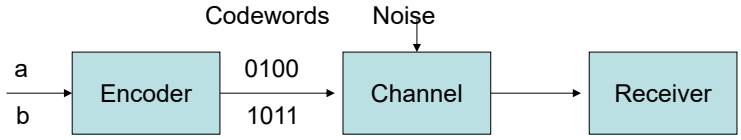


Figure 1.13: Illustration of communications channel in example 1.33.

We construct the probability model with a compound experiment. First, we generate the code word as one of 16 possible binary code words in the sample space

$$\Omega = \{0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111\}$$

The probability measure in this space is given by the signal generation: each atom will have probability $(0.6)^n(0.4)^{4-n}$ where n is the number of ones in the code.

Next, we generate the errors in the code word, based on the error description. Given the code word, the probability that we generate another code word that differs from the current one is $(0.8)^n(0.2)^{4-n}$, where n here refers to the number of bits that were not flipped in error. Hence, $\mathbb{P}[\{0100\}|\{0000\}] = (0.8)^3(0.2)$ We now have a complete probability model, and can answer the following question: If one receives 0010, what is the probability that 0010 was the transmitted message?

$$\mathbb{P}[\{0010\text{transmit}\}|\{0010\}\text{receive}] = \frac{\mathbb{P}[\{0010\text{receive}\}|\{0010\}\text{transmit}]\mathbb{P}[\{0010\text{transmit}\}]}{\mathbb{P}[\{0010\text{receive}\}]}$$

1.4 Computing probability measures for finite sample spaces with equally likely outcomes

There are many cases where one assumes that every outcome is equally likely in an experiment. In this case, the probability of an event $A \in \Omega$ can be computed as the following ratio:

$$\mathbb{P}[A] = \frac{\#\text{ outcomes in } A}{\#\text{ outcomes in } \Omega}.$$

However, we want to avoid enumeration of all the outcomes to figure out the counts! For instance, if we shuffle a deck of 52 cards, what is the number of possible outcomes, where an outcome is a particular order of the 52 cards? Fortunately, we know this is just a problem of permutations, and that number is 52!, which saves us the trouble of enumerating all the card orders! But, consider the event where you are the third player among 4 players in a game of Bridge, and your hand will be dealt all four aces. How many of the shuffle outcomes are in this event? That requires far more clever counting.

Unlike some versions of Probability courses, counting is not a major part of this course. However, the early history of probability up to the 19th century focused on games of chance where counting was the predominant tool for computation of probabilities. In this section, we review some basic formulas that can be used for effective counting.

1.4.1 Counting

The first set of formulas for counting involve permutations and combinations. Given a set of n unique elements, a possible order of these elements is a **permutation**. The number of possible permutations of a set of n elements is $n!$. If we have to select k of these n elements, and order is not important, the number of unique k element sets that can be chosen out of n elements is $\binom{n}{k}$, where

$$\binom{n}{k} = \frac{n!}{(n-k)!(k)!} = \binom{n}{n-k}$$

The number $\binom{n}{k}$ is also called the binomial coefficient. This is because the coefficients in the binomial theorem are given by

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

We will encounter this formula later in this course. For now, it can be used to derive some simple identities such as

$$2^n = \sum_{k=0}^{n} \binom{n}{k},$$

obtained by substituting $a = b = 1$ into the binomial theorem.

Example 1.34
Using the above formulas, we can answer the Bridge question asked earlier: what is the probability that you, sitting as third chair, will be dealt all four aces? We know the number of possible outcomes in Ω, which is the number of permutations of 52 cards: (52)! To compute the number of outcomes where all four aces lie in the cards dealt to the third chair, we proceed as follows:

Assume the four aces are in the cards dealt to the third chair. The number of possible orders for the aces among the 13 cards received by the third chair is $\binom{13}{4}(4!)$, where the first term represents the times you receive an ace, and the second term represents the order in which the aces are received. This number is thus $\frac{13!}{4!9!}4! = (13)(12)(11)(10)$. For each of these possible arrangements of the aces for the third chair, the other 48 cards can be arranged arbitrarily, so there are 48! arrangements of the remaining cards.

We have just computed the number of card shuffles where the third chair will be dealt all four aces as (13)(12)(11)(10) 48!. The probability that this event A happens when all shuffle outcomes are equally likely is:

$$\mathbb{P}[A] = \frac{(13)(12)(11)(10)48!}{52!} = \frac{(13)(12)(11)(10)}{(52)(51)(50)(49)} = \frac{(11)}{(17)(5)(49)} \approx 0.26\%$$

which means that you get all four aces approximately once in every 400 hands (if the dealer is honest.)

Another set of useful counting formulas focus on experiments that are composed of ordered subexperiments. Assume there are r subexperiments, and k^{th} subexperiment consists of n_k outcomes (that can be freely chosen). For instance, you are going to the pet store to buy one of 10 fishes in the small fish tank, one of the 6 dogs in the kennels and one of the 7 cats in the kennel. What are the total number of fish/dog/cat outcomes? The general formula is given by:

$$\#\text{ of outcomes } = n_1 \cdot n_2 \cdots n_r$$

which, for the case of fish/dog/cat outcomes, becomes: # of outcomes = 420.

Counting experiments of this type arise, for instance, in dice games with many dice, or coin games with many coins. If you roll 10 six-sided dice, what are the numbers of possible outcomes? In this case, $n_k = 6$ for each k and $r = 10$, thus it is 6^{10}.

Note that in the above count, we keep track of the outcome for each subexperiment. Thus, order matters in these counts. An implicit assumption is that each subexperiment has outcomes that are selected independently of each other, which is why the total number of outcomes is the product of the number of possible outcomes in each subexperiment.

1.4.2 Sampling

Sampling problems are popular problems in early courses in probability. The typical problem considers a bag with n unique balls (e.g. a lottery urn with 100 numbers). Given that you will take k balls out of the bag, how many possible ways are there to take k balls out? We make a distinction as to whether order matters or not. If order does not matter, this is the simple combination formula we discussed earlier; that is, the number of possible combinations of k balls is $\binom{n}{k}$. However, if order matters, then there are more ways: the right number is $k!\binom{n}{k} = \frac{n!}{(n-k)!}$.

What if the balls are replaced and put back in, so that the same ball can be taken out more than once? We refer to that as sampling with replacement. If order matters, then this is the same as running k subexperiments, each with n possible outcomes, so the total number of outcomes is n^k.

If order does not matter, the number of different outcomes is different. It is hardly obvious as to how to count in this case, as it is not a standard combination. Here is a different way to pose the problem. Assume there are n distinct items. Let $x_i, i = 1, \ldots, n$ denote the number of times an item appears in an outcome; note that this is order-independent. If we are to draw a total of k items, we must have $x_1 + x_2 + \ldots + x_n = k$ in any outcome. Thus, the total number of outcomes is the number of possible solutions of this equation where $x_i \in \{0, 1, \ldots, n\}, i = 1, \ldots, n$.

Let's furthermore represent multiplicities as numbers of ones: Hence, if $x_k = 3$, then $x_k = 111$. Similarly, $x_k = 0$ would be replaced by $x_k =$, that is, no entry. With this notation, the term $x_1 + x_2 + \ldots + x_n$ must be a sequence of length $k + (n-1)$ composed of exactly k digits 1, and $n - 1$ + signs! This is the hard part to visualize: we have reduced the problem to finding $n - 1$ positions for the + signs out of the $k + n - 1$ total positions. For instance, if $n = 3$ and $k = 2$, the sequence ++11 means $x_1 = x_2 = 0, x_3 = 2$. The sequence 1+1+ means $x_1 = x_2 = 1, x_3 = 0$. Once you understand this mapping, the final answer is just another combination formula:

$$\#\text{ order-independent } k \text{ out of } n \text{ samples with replacement} = \binom{n-1+k}{n-1} = \binom{n-1+k}{k}$$

Example 1.35
Suppose we have three balls, one red, one blue and one green. We put them in a bag, and sample them three times with replacement. How many order-independent sets of colors can be obtained from this sampling?

In this example, the number of draws $k = 3$, and the number of possible colors (values) is $n = 3$. Hence, the total number of order-independent outcomes is

$$\binom{n-1+k}{k} = \binom{3-1+3}{3} = \binom{5}{3} = 10.$$

For this small example, we can actually list all of the possible outcomes. We list outcome R for a red ball, G for a green ball, and B for a blue ball. The outcomes are

RRR RRB RRG RBB RBG RGG BBB BBG BGG GGG

We summarize these formulas in Table 1.1:

	Order Dependent	Order Independent
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Table 1.1: Sampling formulas

1.4.3 Partitions

Another popular set of examples in elementary probability courses consist of partition problems where we have n items and we want to divide them into r groups, so that the k^{th} group contains n_k elements, such that $n_1 + n_2 + \ldots + n_r = n$. For instance, we have 18 basketball players, and we want to divide them into six teams of three players. How many possible ways of forming teams are there?

This is an extension of the binomial coefficient formula. In that problem we wanted to divide n elements into two groups, one of size k and another of size $n-k$. In this extension each element appears in exactly one group because $\sum_{k=1}^{r} n_k = n$. The number of ways to form such a **partition** is given by the multinomial coefficient

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

The way to derive this is to perform sequential selection of combinations: First, select n_1 out of n. Then select n_2 out of the remaining $n - n_1$. Continue this until you select n_r out of the remaining n_r. This yields:

$$\begin{aligned}
\binom{n}{n_1, n_2, \ldots, n_r} &= \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_r}{n_r} \\
&= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!}\cdots\frac{n_r!}{n_r!} \\
&= \frac{n!}{n_1! n_2! \cdots n_r!} \text{ because of all the cancelations with successive terms.}
\end{aligned}$$

This formula can be used in a generalization of the binomial theorem, as follows:

$$(x_1 + x_2 + x_3 + \ldots + x_r)^n = \sum_{n_1+n_2+\ldots+n_r=n} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

which lead to some nice identities such as

$$r^n = \sum_{n_1+n_2+\ldots+n_r=n} \frac{n!}{n_1!n_2!\cdots n_r!}.$$

Example 1.36
Twelve people have a potluck party. Six people will be selected to bring a main dish, four people will bring drinks, and two people will bring dessert. How many ways can they be divided into these three groups?

We solve this using the multinomial partition formula:

$$\#\text{ partitions} = \binom{12}{6,4,2} = \frac{12!}{6!48} = \frac{(12)(11)(10)(9)(8)(7)}{48} = (22)(10)(9)(7) = 13,860.$$

Example 1.37
Suppose we have a lottery with numbers from 1 to 59. You are allowed to select five numbers, and you can choose the same number more than once. What is the probability of winning? Note that the numbers you select must be picked in the same order as the lottery to win.

What are the number of ways that five numbers can be chosen by the lottery? This is order-dependent sampling with replacement, so the formula is:

$$59^5 = 714,924,299.$$

Your chance of winning is one in 714,924,299.

If we only allowed selection of a number once (instead of putting the selected number back in the lottery urn), we would be doing order-dependent sampling without replacement, so the answer is

$$\frac{59!}{54!} = \frac{59 \times 58 \times 57 \times 56 \times 55}{=} 600,766,320$$

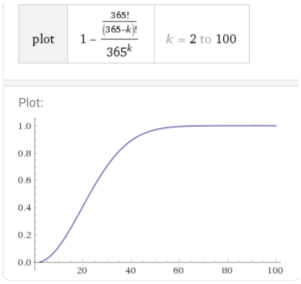
which increases your chances of winning a little bit.

Example 1.38
Assume you have a perfectly shuffled deck of cards. If you draw five cards, without replacement, what is the probability that exactly three of the five are kings? Note that this is not order-dependent. To answer this, we compute the number of five card hands with exactly three kings. The three kings must be chosen from four possible kings, and the other two cards chosen from 48 non-king cards. This gives the number of hands with three kings as $\binom{4}{3}\binom{48}{2}$. The total number of five card hands, with replacement, is $\binom{52}{5}$. Then,

$$\mathbb{P}[\{\text{Choose exactly 3 kings in five cards}\}] = \frac{\frac{4\cdot48\cdot47}{2}}{\frac{52\cdot51\cdot50\cdot49\cdot48}{5\cdot4\cdot3\cdot2}} = \frac{4512}{2598960} = 0.001736.$$

Here is another common example used to surprise a class.

Example 1.39
In a class with k students, assuming that each student was equally likely to be born in one of the 365 days in a regular calendar year, what is the probability that two or more students share a birthday?
Sometimes, it is easier to compute the probability of the complement of an event. In this case, we compute the probability that no student shares a common birthday. The number of ways to select k birthdays uniquely are $\frac{365!}{(365-k)!}$. The total number of ways to select k birthdays is 365^k. Hence, the probability that k students do not have a birthday in common among any two of them is



$$\mathbb{P}[\{\text{No common birthday in } k \text{ students}\}] = \frac{\frac{365!}{(365-k)!}}{365^k}.$$

Then, the probability that at least two students share a common birthday is

$$\mathbb{P}[\{\text{At least two students share a common birthday among } k \text{ students}\}] = 1 - \frac{\frac{365!}{(365-k)!}}{365^k}.$$

Note that this quickly approaches one! For $k = 64$ this is approximately 0.997. Our class size is bigger. For $k = 100$, the probability is 0.9999997, nearly 1. The curve of how the probability grows with k is shown in Figure 1.39.

1.4.4 Independent Trials

Up to now, we have assumed that every outcome was equally likely, and therefore we would compute probabilities by counting the number of outcomes. Jacob Bernoulli developed an extension for this were an experiment consisted of multiple identical, independent subexperiments, each with two possible outcomes (e.g. win or lose). However, the probability of an outcome in each subexperiment was unbalanced: the probability of winning was different than the probability of losing.

We refer to these subexperiments as Bernoulli trials. A Bernoulli trial is a random experiment with two outcomes, say "win" and "lose", where we define the probability of "win" as a number $p \in [0,1]$ and the probability of "lose" as $1-p$. For instance, a Bernoulli trial might be the outcome of a coin toss, where "heads" corresponds to "win".

If an experiment performs n independent Bernoulli trials, how many possible ways are there to get a total of k "win" outcomes? This is a combination problem, were we are going to select the k positions that have a "win" outcome among the n trials, and that is $\binom{n}{k}$. However, what is the probability of each of those combined outcomes? In each of those combined outcomes there are k subexperiments that resulted in "win" and $n-k$ subexperiments that resulted in "lose". Since these are independent subexperiments the probability of each of those outcomes is $p^k(1-p)^{n-k}$, as the probability for the combined outcome is the product of the probabilities of each subexperiment. Hence, the probability of the event $\{k$ "win" outcomes in n Bernoulli trials $\}$ is the sum of the probabilities of each outcome in the event, which is

$$\mathbb{P}[\{k \text{ "win" outcomes in } n \text{ Bernoulli trials }\}] = \binom{n}{k} p^k (1-p)^{n-k}$$

This distribution, discovered by Bernoulli, is termed the binomial distribution.

Example 1.40
You have a biased coin, such that "heads" occurs with probability 0.6. If you flip the coin 10 times, what is the probability that you have a total of 4 "heads" outcomes? Applying the above formula yields

$$\mathbb{P}[\{4 \text{ "heads" out of } 10\}] = \binom{10}{4} 0.6^4 0.4^6 \approx 0.1115.$$

Let's generalize the above result. Suppose that, instead of having two outcomes each subexperiment has r possible outcomes $a_1, \ldots, a_r$ with probabilities $p_1, \ldots, p_r$. We want to conduct n independent subexperiments and count the number of outcomes of each type. That is, we want to count the total number of outcomes n_1 of a_1, the number of outcomes n_2 of $a_2, \ldots$, and the total number of outcomes n_r of a_r. We know that the total number of partitions of n outcomes into r classes with n_k outcomes of class k is $\binom{n}{n_1, n_2, \ldots, n_r}$. What is important is that each of those partitions has the same probability because of the independence of the subexperiments: $p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$. This yields the following formula for the probability that, when we run n subexperiments, we will get n_1 outcomes of a_1, n_2 of a_2, $\ldots$, n_r of a_r:

$$\mathbb{P}\big[\{n_1 \text{ occurrences of } a_1, \ldots, n_r \text{ occurrences of } a_r\}\big] = \binom{n}{n_1, n_2, \ldots, n_r} p_1^{n_1} \cdots p_r^{n_r}$$

Example 1.41
I have a game with three outcomes: win, lose, draw. The probability of win is 0.4, lose 0.5, draw 0.1. If I play 10 times with independent outcomes, what is the probability of 4 wins, 4 lose and 2 draw?

$$\mathbb{P}[\text{win 4, lose 4, draw 2 of 10}] = \binom{10}{4,4,2} 0.4^4 0.5^4 0.1^2 = \frac{10!}{4!4!2!} \frac{1}{(2500)(25)} = \frac{(10)(9)(8)(7)(6)(5)}{(48)(2500)(25)} = \frac{(9)(7)}{(50)(25)} = 0.0504$$

Example 1.42
In a wireless channel, bits are flipped independently in error with probability 0.01. If you transmit 100 bits, what is the probability that 3 of them are flipped?

$$\mathbb{P}[\{\text{ 3 of 100 bits are flipped}\}] = \binom{100}{3}(0.01)^3(0.99)^{97} \approx 0.061.$$

For the same example, if we have an error correcting code that can correct up to three bits in error, what is the probability that all bits are recovered without error? This is the probability that no bits are flipped, plus the probability that one bit is flipped, plus the probability that two bits are flipped. That is,

$$\begin{aligned}\mathbb{P}[\{\text{ All bits recovered}\}] &= \binom{100}{0}(0.99)^{100} + \binom{100}{1}(0.01)(0.99)^{99} + \binom{100}{2}(0.01)^2(0.99)^{98} + \binom{100}{3}(0.01)^3(0.99)^{97} \\ &\approx 0.982.\end{aligned}$$