Undergraduate Probability I

Class Notes for
Engineering Students

Free Research and Education Documents

Fall 2011

Contents

Chapter 1

Mathematical Review

Set theory is now generally accepted as the foundation of modern mathematics, and it plays an instrumental role in the treatment of probability. Unfortunately, a simple description of set theory can lead to paradoxes, while a rigorous axiomatic approach is quite tedious. In these notes, we circumvent these difficulties and assume that the meaning of a set as a collection of objects is intuitively clear. This standpoint is known as naive set theory. We proceed to define relevant notation and operations.

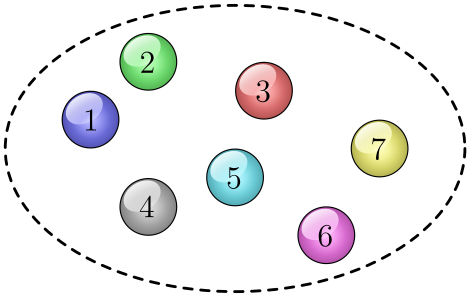

Figure 1.1: This is an illustration of a generic set and its elements.

A *set* is a collection of objects, which are called the *elements* of the set. If an element x belongs to a set S, we express this fact by writing $x \in S$. If x does not belong to S, we write $x \notin S$. We use the equality symbol to denote *logical identity*. For instance, $x = y$ means that x and y are symbols denoting the same object. Similarly, the equation $S = T$ states that S and T are two symbols for the same set. In particular, the sets S and T contain precisely the same elements. If x and y are different objects then we write $x \neq y$. Also, we can express the fact that S and T are different sets by writing $S \neq T$.

A set S is a *subset* of T if every element of S is also contained in T. We express this relation by writing $S \subset T$. Note that this definition does not require S to be different from T. If $S \subset T$ and S is different from T, then S is a *proper subset* of T, which we indicate by $S \subsetneq T$. Moreover, $S = T$ if and only if $S \subset T$ and $T \subset S$. In fact, this latter statement outlines a methodology to show that two sets are equal.

There are many ways to specify a set. If the set contains only a few elements, one can simply list the objects in the set;

$$S = \{x_1, x_2, x_3\}.$$

The content of a set can also be enumerated whenever S has a countable number of elements,

$$S = \{x_1, x_2, \ldots\}.$$

Usually, the way to specify a set is to take some collection T of objects and some property that elements of T may or may not possess, and to form the set consisting of all elements of T having that property. For example, starting with the integers $\mathbb{Z}$, we can form the subset S consisting of all even numbers,

$$S = \{x \in \mathbb{Z} | x \text{ is an even number}\}.$$

More generally, we denote the set of all elements that satisfy a certain property P by $S = \{x | x \text{ satisfies } P\}$. The braces are to be read as “the set of” while the symbol $|$ stands for the words “such that.”

It is convenient to introduce two special sets. The *empty set*, denoted by $\emptyset$, is a set that contains no elements. The *universal set* is the collection of all objects of interest in a particular context, and it is represented by Ω. Once a universal set Ω is specified, we need only consider sets that are subsets of Ω. In the context of probability, Ω is often called the *sample space*. The *complement* of a set S, relative to the universal set Ω, is the collection of all objects in Ω that do not belong to S,

$$S^{\mathrm{c}} = \{x \in \Omega | x \notin S\}.$$

For example, we have $\Omega^{\mathrm{c}} = \emptyset$.

1.1 Elementary Set Operations

The field of probability makes extensive use of set theory. Below, we review elementary set operations, and we establish basic terminology and notation. Consider two sets, S and T.

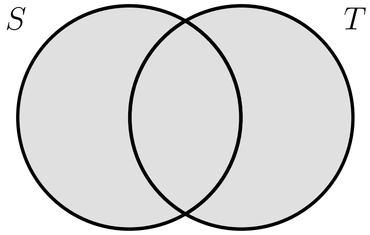

Figure 1.2: This is an abstract representation of two sets, S and T. Each contains the elements in its shaded circle and the overlap corresponds to elements that are in both sets.

The *union* of sets S and T is the collection of all elements that belong to S or T (or both), and it is denoted by $S \cup T$. Formally, we define the union of these two sets by $S \cup T = \{x | x \in S \text{ or } x \in T\}$.

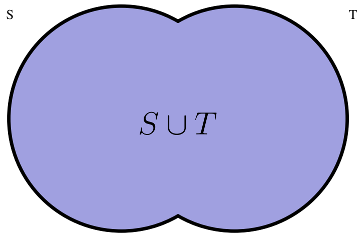

Figure 1.3: The union of sets S and T consists of all elements that are contained in S or T.

The *intersection* of sets S and T is the collection of all elements that belong to both S and T. It is denoted by $S \cap T$, and it can be expressed mathematically as $S \cap T = \{x | x \in S \text{ and } x \in T\}$.

When S and T have no elements in common, we write $S \cap T = \emptyset$. We also express this fact by saying that S and T are *disjoint*. More generally, a collection of sets is said to be disjoint if no two sets have a common element.

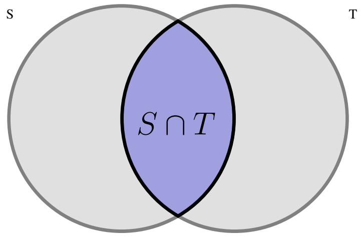

Figure 1.4: The intersection of sets S and T only contains elements that are both in S and T.

A collection of sets is said to form a *partition* of S if the sets in the collection are disjoint and their union is S.

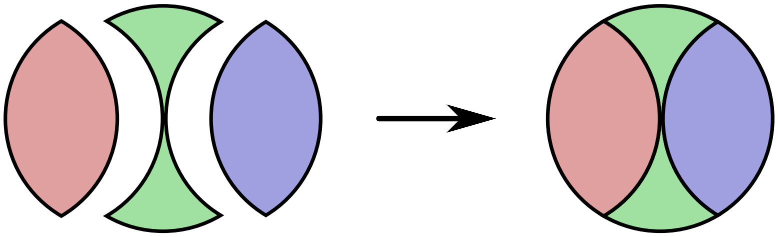

Figure 1.5: A partition of S is a collection of sets that are disjoint and whose union is S.

The *difference* of two sets, denoted by $S-T$, is defined as the set consisting of those elements of S that are not in T, $S - T = \{x|x \in S \text{ and } x \notin T\}$. This set is sometimes called the complement of T relative to S, or the complement of T in S.

So far, we have looked at the definition of the union and the intersection of two sets. We can also form the union or the intersection of arbitrarily many sets. This is defined in a straightforward way,

$$\begin{aligned}\bigcup_{i \in \mathbb{I}} S_i &= \{x|x \in S_i \text{ for some } i \in \mathbb{I}\}\\ \bigcap_{i \in \mathbb{I}} S_i &= \{x|x \in S_i \text{ for all } i \in \mathbb{I}\}.\end{aligned}$$

The index set $\mathbb{I}$ can be finite or infinite.

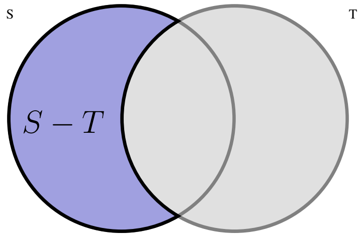

Figure 1.6: The complement of T relative to S contains all the elements of S that are not in T.

1.2 Additional Rules and Properties

Given a collection of sets, it is possible to form new ones by applying elementary set operations to them. As in algebra, one uses parentheses to indicate precedence. For instance, $R \cup (S \cap T)$ denotes the union of two sets R and $S \cap T$, whereas $(R \cup S) \cap T$ represents the intersection of two sets $R \cup S$ and T. The sets thus formed are quite different.

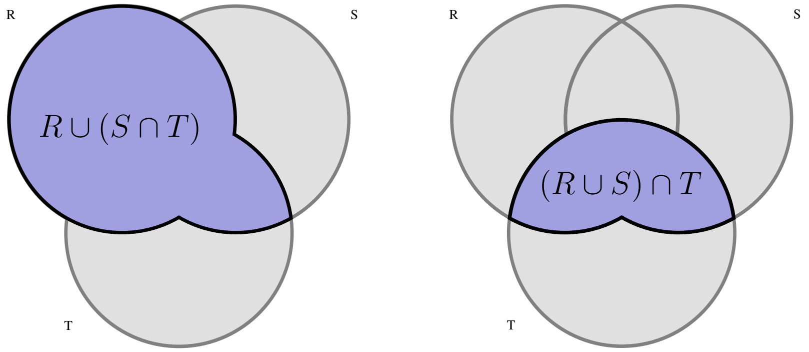

Figure 1.7: The order of set operations is important; parentheses should be employed to specify precedence.

Sometimes, different combinations of operations lead to a same set. For instance, we have the following distributive laws

$$R \cap (S \cup T) = (R \cap S) \cup (R \cap T)$$
$$R \cup (S \cap T) = (R \cup S) \cap (R \cup T).$$

Two particularly useful equivalent combinations of operations are given by

De Morgan's laws, which state that

$$\begin{aligned} R - (S \cup T) &= (R - S) \cap (R - T) \\ R - (S \cap T) &= (R - S) \cup (R - T). \end{aligned}$$

These two laws can also appear in different forms,

$$\begin{aligned} \left(\bigcup_{i \in \mathbb{I}} S_i \right)^{\mathrm{c}} &= \bigcap_{i \in \mathbb{I}} S_i^{\mathrm{c}} \\ \left(\bigcap_{i \in \mathbb{I}} S_i \right)^{\mathrm{c}} &= \bigcup_{i \in \mathbb{I}} S_i^{\mathrm{c}} \end{aligned}$$

when multiple sets are involved. To establish the first equality, suppose that x belongs to $\left(\bigcup_{i \in \mathbb{I}} S_i \right)^{\mathrm{c}}$. Then x is not contained in $\bigcup_{i \in \mathbb{I}} S_i$. That is, x is not an element of S_i for any $i \in \mathbb{I}$. This implies that x belongs to S_i^{c} for all $i \in \mathbb{I}$, and therefore $x \in \bigcap_{i \in \mathbb{I}} S_i^{\mathrm{c}}$. We have shown that $\left(\bigcup_{i \in \mathbb{I}} S_i \right)^{\mathrm{c}} \subset \bigcap_{i \in \mathbb{I}} S_i^{\mathrm{c}}$. The converse inclusion is obtained by reversing the above argument. The second law can be obtained in a similar fashion.

1.3 Cartesian Products

There is yet another way to create new sets from existing ones. It involves the notion of an *ordered pair* of objects. Given sets S and T, the *cartesian product* $S \times T$ is the set of all ordered pairs (x, y) for which x is an element of S and y is an element of T, $S \times T = \{(x, y) | x \in S \text{ and } y \in T\}$.

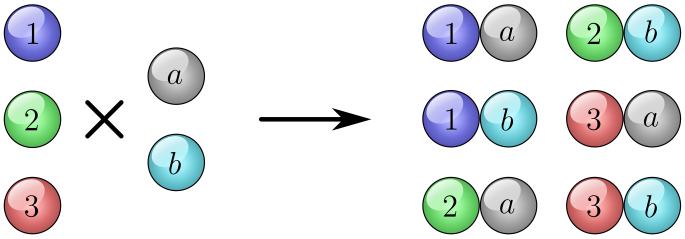

Figure 1.8: Cartesian products can be used to create new sets. In this example, the sets $\{1, 2, 3\}$ and $\{a, b\}$ are employed to create a cartesian product with six elements.

1.4 Functions

A *function* is a special type of relation that assigns exactly one value to each input. A common representation for a function is $f(x) = y$, where f denotes the rule of correspondence. The *domain* of a function is the set over which this function is defined; that is, the collection of arguments that can be used as input. The *codomain* is the set into which all the outputs of the function are constrained to lie. To specify these sets explicitly, the notation

$$f : X \to Y$$

is frequently used, where X indicates the domain and Y is the codomain. In these notes, we adopt an intuitive point of view, with the understanding that a function maps every argument to a unique value in the codomain. However, a function can be defined formally as a triple (X, Y, F) where F is a structured subset of the Cartesian product $X \times Y$.

Example 1. *Consider the function $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = x^2$. In this case, the domain $\mathbb{R}$ and the codomain $\mathbb{R}$ are identical. The rule of correspondence for this function is $x \mapsto x^2$, which should be read "x maps to x^2."*

Example 2. *An interesting function that plays an important role in probability is the indicator function. Suppose S is a subset of the real numbers. We define the indicator function of set S, denoted $\mathbf{1}_S : \mathbb{R} \to \{0, 1\}$, by*

$$\mathbf{1}_S(x) = \begin{cases} 1, & x \in S \\ 0, & x \notin S. \end{cases}$$

In words, the value of the function $\mathbf{1}_S(\cdot)$ indicates whether its argument belongs to S or not. A value of one represents inclusion of the argument in S, whereas a zero signifies exclusion.

The *image* of function is the set of all objects of the form $f(x)$, where x ranges over the elements of X,

$$\{f(x) \in Y | x \in X\}.$$

The image of $f : X \to Y$ is sometimes denoted by $f(X)$ and it is, in general, a subset of the codomain. Similarly, the *preimage* of a set $T \subset Y$ under $f : X \to Y$ is the subset of X defined by

$$f^{-1}(T) = \{x \in X | f(x) \in T\}.$$

It may be instructive to point out that the preimage of a singleton set can contain any number of elements,

$$f^{-1}(\{y\}) = \{x \in X | f(x) = y\}.$$

This set, $f^{-1}(\{y\})$, is sometimes called the *level set* of y.

A function is *injective* or *one-to-one* if it preserves distinctness; that is, different elements from the domain never map to a same element in the codomain. Mathematically, the function $f : X \to Y$ is injective if $f(x_1) = f(x_2)$ implies $x_1 = x_2$ for all $x_1, x_2 \in X$. The function f is *surjective* or *onto* if its image is equal to its codomain. More specifically, a function $f : X \to Y$ is surjective if and only if, for every $y \in Y$, there exists $x \in X$ such that $f(x) = y$. Finally, a function that is both one-to-one and onto is called a *bijection*. A function f is bijective if and only if its inverse relation f^{-1} is itself a function. In this case, the preimage of a singleton set is necessarily a singleton set. The inverse of f is then represented unambiguously by $f^{-1} : Y \to X$, with $f^{-1}(f(x)) = x$ and $f(f^{-1}(y)) = y$.

1.5 Set Theory and Probability

Set theory provides a rigorous foundation for modern probability and its axiomatic basis. It is employed to describe the laws of probability, give meaning to their implications and answer practical questions. Becoming familiar with basic definitions and set operations is key in understanding the subtleties of probability; it will help overcome its many challenges. A working knowledge of set theory is especially critical when modeling measured quantities and evolving processes that appear random, an invaluable skill for engineers.

Further Reading

1. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 1.1.

2. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Section 1.2.

Chapter 2

Combinatorics and Intuitive Probability

The simplest probabilistic scenario is perhaps one where the set of possible outcomes is finite and these outcomes are all equally likely. In such cases, computing the probability of an event amounts to counting the number of elements comprising this event and then dividing the sum by the total number of admissible outcomes.

Example 3. *The rolling of a fair die is an experiment with a finite number of equally likely outcomes, namely the different faces labeled one through six. The probability of observing a specific face is equal to*

$$\frac{1}{\text{\textit{Number of faces}}} = \frac{1}{6}.$$

Similarly, the probability of an arbitrary event can be computed by counting the number of distinct outcomes included in the event. For instance, the probability of rolling a prime number is

$$\Pr(\{2, 3, 5\}) = \frac{\text{\textit{Number of outcomes in event}}}{\text{\textit{Total number of outcomes}}} = \frac{3}{6}.$$

While counting outcomes may appear intuitively straightforward, it is in many circumstances a daunting task. Calculating the number of ways that certain patterns can be formed is part of the field of *combinatorics*. In this chapter, we introduce useful counting techniques that can be applied to situations pertinent to probability.

2.1 The Counting Principle

The counting principle is a guiding rule for computing the number of elements in a cartesian product. Suppose that S and T are finite sets with m and n elements, respectively. The cartesian product of S and T is given by

$$S \times T = \{(x, y) | x \in S \text{ and } y \in T\}.$$

The number of elements in the cartesian product $S \times T$ is equal to mn. This is illustrated in Figure 2.1.

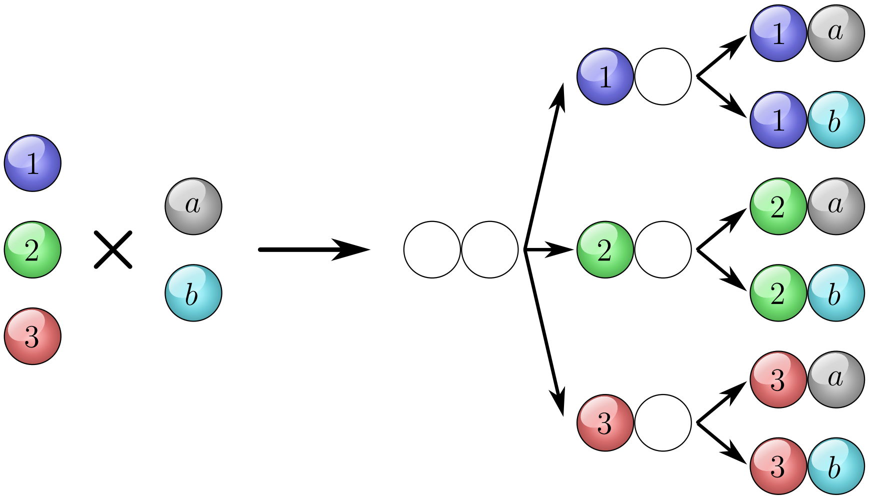

Figure 2.1: This figure provides a graphical interpretation of the cartesian product of $S = \{1, 2, 3\}$ and $T = \{a, b\}$. In general, if S has m elements and T contains n elements, then the cartesian product $S \times T$ consists of mn elements.

Example 4. *Consider an experiment consisting of flipping a coin and rolling a die. There are two possibilities for the coin, heads or tails, and the die has six faces. The number of possible outcomes for this experiment is* $2 \times 6 = 12$. *That is, there are twelve different ways to flip a coin and roll a die.*

The counting principle can be broadened to calculating the number of elements in the cartesian product of multiple sets. Consider the finite sets $S_1, S_2, \ldots, S_r$ and their cartesian product

$$S_1 \times S_2 \times \cdots \times S_r = \{(s_1, s_2, \ldots, s_r) | s_i \in S_i\}.$$

If we denote the cardinality of S_i by $n_i = |S_i|$, then the number of distinct ordered r-tuples of the form $(s_1, s_2, \ldots, s_r)$ is $n = n_1 n_2 \cdots n_r$.

Example 5 (Sampling with Replacement and Ordering)**.** *An urn contains n balls numbered one through n. A ball is drawn from the urn and its number is recorded on an ordered list. The ball is then replaced in the urn. This procedure is repeated k times. We wish to compute the number of possible sequences that can result from this experiment. There are k drawings and n possibilities per drawing. Using the counting principle, we gather that the number of distinct sequences is n^k.*

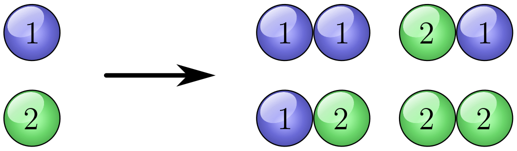

Figure 2.2: The cartesian product $\{1, 2\}^2$ has four distinct ordered pairs.

Example 6. *The* power set *of S, denoted by 2^S, is the collection of all subsets of S. In set theory, 2^S represents the set of all functions from S to $\{0, 1\}$. By identifying a function in 2^S with the corresponding preimage of one, we obtain a bijection between 2^S and the subsets of S. In particular, each function in 2^S is the characteristic function of a subset of S.*

Suppose that S is finite with $n = |S|$ elements. For every element of S, a characteristic function in 2^S is either zero or one. There are therefore 2^n distinct characteristic functions from S to $\{0, 1\}$. Hence, the number of distinct subsets of S is given by 2^n.

2.2 Permutations

Again, consider the integer set $S = \{1, 2, \ldots, n\}$. A *permutation* of S is an ordered arrangement of its elements, i.e., a list without repetitions. The number of permutations of S can be computed as follows. Clearly, there are n distinct possibilities for the first item in the list. The number of possibilities for the second item is $n - 1$, namely all the integers in S except the element

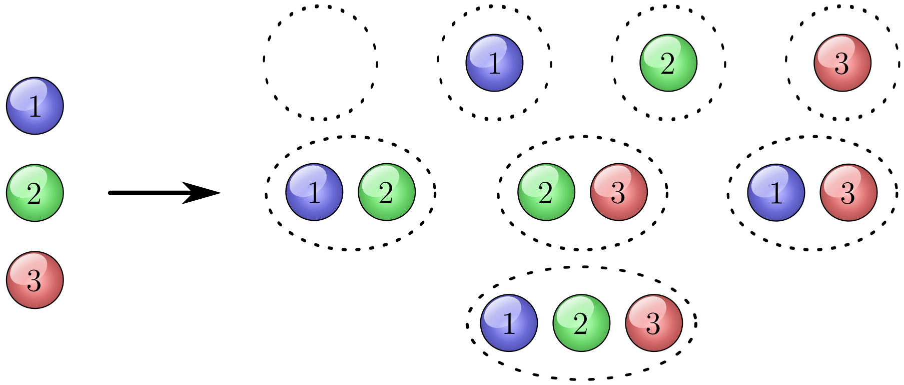

Figure 2.3: The power set of $\{1, 2, 3\}$ contains eight subsets. These elements are displayed above.

we selected initially. Similarly, the number of distinct possibilities for the mth item is $n - m + 1$. This pattern continues until all the elements in S are recorded. Summarizing, we find that the total number of permutations of S is n *factorial*, $n! = n(n-1) \cdots 1$.

Example 7. *We wish to compute the number of permutations of* $S = \{1, 2, 3\}$. *Since the set* S *possesses three elements, it has* $3! = 6$ *different permutations. They can be written as* $123, 132, 213, 231, 312, 321$.

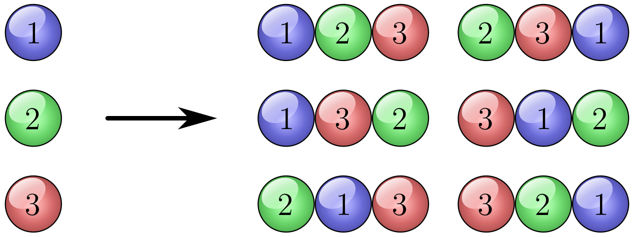

Figure 2.4: Ordering the numbers one, two and three leads to six possible permutations.

2.2.1 Stirling's Formula*

The number $n!$ grows very rapidly as a function of n. A good approximation for $n!$ when n is large is given by *Stirling's formula*,

$$n! \sim n^n e^{-n} \sqrt{2\pi n}.$$

The notation $a_n \sim b_n$ signifies that the ratio $a_n/b_n \to 1$ as $n \to \infty$.

2.2.2 k-Permutations

Suppose that we rank only k elements out of the set $S = \{1, 2, \ldots, n\}$, where $k \leq n$. We wish to count the number of distinct k-permutations of S. Following our previous argument, we can choose one of n elements to be the first item listed, one of the remaining $(n-1)$ elements for the second item, and so on. The procedure terminates when k items have been recorded. The number of possible sequences is therefore given by

$$\frac{n!}{(n-k)!} = n(n-1) \cdots (n-k+1).$$

Example 8. *A recently formed music group can play four original songs. They are asked to perform two songs at South by Southwest. We wish to compute the number of song arrangements the group can offer in concert. Abstractly, this is equivalent to computing the number of 2-permutations of four songs. Thus, the number of distinct arrangements is* $4!/2! = 12$.

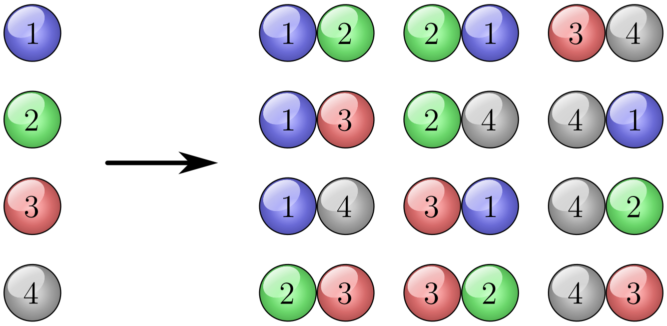

Figure 2.5: There are twelve 2-permutations of the numbers one through four.

Example 9 (Sampling without Replacement, with Ordering). *An urn contains n balls numbered one through n. A ball is picked from the urn, and its number is recorded on an ordered list. The ball is not replaced in the urn. This procedure is repeated until k balls are selected from the urn, where $k \leq n$. We wish to compute the number of possible sequences that can result from this experiment. The number of possibilities is equivalent to the number of k-permutations of n elements, which is given by $n!/(n-k)!$.*

2.3 Combinations

Consider the integer set $S = \{1, 2, \ldots, n\}$. A *combination* is a subset of S. We emphasize that a combination differs from a permutation in that elements in a combination have no specific ordering. The 2-element subsets of $S = \{1, 2, 3, 4\}$ are

$$\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\},$$

whereas the 2-permutations of S are more numerous with

$$\begin{gathered}(1,2), (1,3), (1,4), (2,1), (2,3), (2,4), \\ (3,1), (3,2), (3,4), (4,1), (4,2), (4,3).\end{gathered}$$

Consequently, there are fewer 2-element subsets of S than 2-permutations of S.

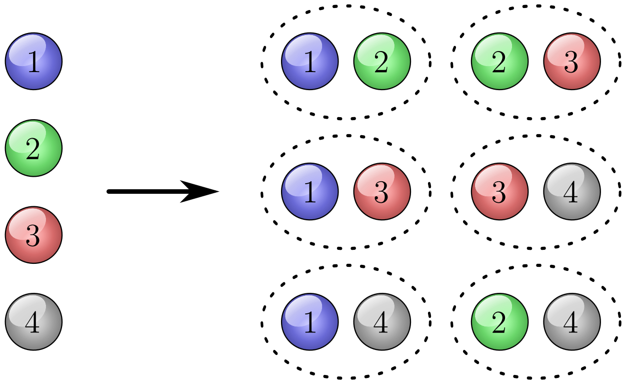

Figure 2.6: There exist six 2-element subsets of the numbers one through four.

We can compute the number of k-element combinations of $S = \{1, 2, \ldots, n\}$ as follows. Note that a k-permutation can be formed by first selecting k objects from S and then ordering them. There are $k!$ distinct ways of ordering k components. The number of k-permutations must therefore be equal to the number of k-element combinations multiplied by $k!$. Since the total number of k-permutations of S is $n!/(n-k)!$, we gather that the number of k-element combinations is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

This expression is termed a *binomial coefficient*. Observe that selecting a k-element subset of S is equivalent to choosing the $n-k$ elements that belong to its complement. Thus, we can write

$$\binom{n}{k} = \binom{n}{n-k}.$$

Example 10 (Sampling without Replacement or Ordering)**.** *An urn contains n balls numbered one through n. A ball is drawn from the urn and placed in a separate jar. This process is repeated until the jar contains k balls, where $k \leq n$. We wish to compute the number of distinct combinations the jar can hold after the completion of this experiment. Because there is no ordering in the jar, this amounts to counting the number of k-element subsets of a given n-element set, which is given by*

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

Again, let $S = \{1, 2, \ldots, n\}$. Since a combination is also a subset and the number of k-element combinations of S is $\binom{n}{k}$, the sum of the binomial coefficients $\binom{n}{k}$ over all values of k must be equal to the number of elements in the power set of S,

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

2.4 Partitions

Abstractly, a combination is equivalent to partitioning a set into two disjoint subsets, one containing k objects and the other containing the $n-k$ remaining elements. In general, the set $S = \{1, 2, \ldots, n\}$ can be partitioned into r disjoint subsets. Let $n_1, n_2, \ldots, n_r$ be nonnegative integers such that

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

Consider the following iterative algorithm that leads to a *partition* of S. First, we choose a subset of n_1 elements from S. Having selected the first subset, we pick a second subset containing n_2 elements from the remaining $n - n_1$ elements. We continue this procedure by successively choosing subsets of n_i elements from the residual $n - n_1 - \cdots - n_{i-1}$ elements, until no element remains. This algorithm yields a partition of S into r subsets, with the ith subset containing exactly n_i elements.

We wish to count the number of such partitions. We know that there are $\binom{n}{n_1}$ ways to form the first subset. Examining our algorithm, we see that there are exactly

$$\binom{n - n_1 - \cdots - n_{i-1}}{n_i}$$

ways to form the ith subset. Using the counting principle, the total number of partitions is then given by

$$\binom{n}{n_1}\binom{n - n_1}{n_2} \cdots \binom{n - n_1 - \cdots - n_{r-1}}{n_r},$$

which after simplification can be written as

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

This expression is called a *multinomial coefficient.*

Example 11. *A die is rolled nine times. We wish to compute the number of possible outcomes for which every odd number appears three times. The number of distinct sequences in which one, three and five each appear three times is equal to the number of partitions of* $\{1, 2, \ldots, 9\}$ *into three subsets of size three, namely*

$$\frac{9!}{3!3!3!} = 1680.$$

In the above analysis, we assume that the cardinality of each subset is fixed. Suppose instead that we are interested in counting the number of ways to pick the cardinality of the subsets that form the partition. Specifically, we wish to compute the number of ways integers $n_1, n_2, \ldots, n_r$ can be selected such that every integer is nonnegative and

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

We can visualize the number of possible assignments as follows. Picture n balls spaced out on a straight line and consider $r-1$ vertical markers, each of which can be put between two consecutive balls, before the first ball, or after the last ball. For instance, if there are five balls and two markers then one possible assignment is illustrated in Figure 2.7.

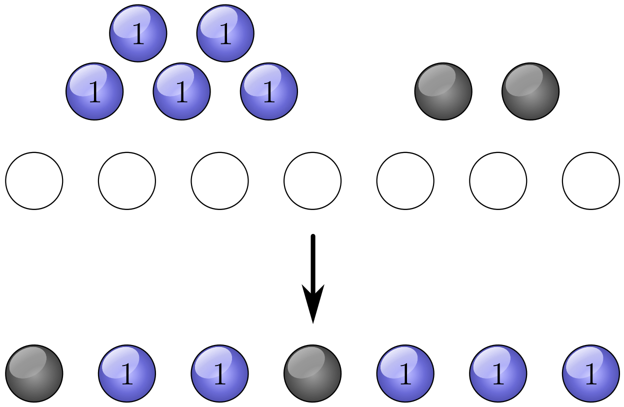

Figure 2.7: The number of possible cardinality assignments for the partition of a set of n elements into r distinct subsets is equivalent to the number of ways to select n positions out of $n+r-1$ candidates.

The number of objects in the first subset corresponds to the number of balls appearing before the first marker. Similarly, the number of objects in the ith subset is equal to the number of balls positioned between the ith marker and the preceding one. Finally, the number of objects in the last subset is simply the number of balls after the last marker. In this scheme, two consecutive markers imply that the corresponding subset is empty. For example, the integer assignment associated with the Figure 2.7 is

$$(n_1, n_2, n_3) = (0, 2, 3).$$

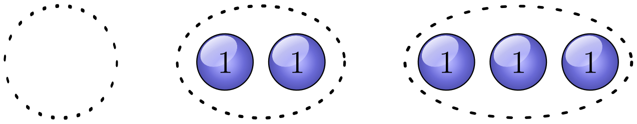

Figure 2.8: The assignment displayed in Figure 2.7 corresponds to having no element in the first set, two elements in the second set and three elements in the last one.

There is a natural relation between an integer assignment and the graphical representation depicted above. To count the number of possible integer assignments, it suffices to calculate the number of ways to place the markers and the balls. In particular, there are $n + r - 1$ positions, n balls and $r - 1$ markers. The number of ways to assign the markers is equal to the number of n-combinations of $n + r - 1$ elements,

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

Example 12 (Sampling with Replacement, without Ordering)**.** *An urn contains r balls numbered one through r. A ball is drawn from the urn and its number is recorded. The ball is then returned to the urn. This procedure is repeated a total of n times. The outcome of this experiment is a table that contains the number of times each ball has come in sight. We are interested in computing the number of possible outcomes. This is equivalent to counting the ways a set with n elements can be partitioned into r subsets. The number of possible outcomes is therefore given by*

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

2.5 Combinatorial Examples

In this section, we present a few applications of combinatorics to computing the probabilities of specific events.

Example 13 (Pick 3 Texas Lottery)**.** *The Texas Lottery game "Pick 3" is easy to play. A player must pick three numbers from zero to nine, and choose how to play them: exact order or any order. The Pick 3 balls are drawn using three air-driven machines. These machines employ compressed air to mix and select each ball.*

The probability of winning when playing the exact order is

$$\frac{1}{10}\frac{1}{10}\frac{1}{10} = \frac{1}{1000}.$$

The probability of winning while playing any order depends on the numbers selected. When three distinct numbers are selected, the probability of winning is given by

$$\frac{3!}{1000} = \frac{3}{500}.$$

If a number is repeated twice, the probability of winning becomes

$$\frac{\binom{3}{2}}{1000} = \frac{3}{1000}.$$

Finally, if a same number is selected three times, the probability of winning decreases to 1/1000.

Example 14 (Mega Millions Texas Lottery)**.** *To play the Mega Millions game, a player must select five numbers from 1 to 56 in the upper white play area of the play board, and one Mega Ball number from 1 to 46 in the lower yellow play area of the play board. All drawing equipment is stored in a secured on-site storage room. Only authorized drawings department personnel have keys to the door. Upon entry of the secured room to begin the drawing process, a lottery drawing specialist examines the security seal to determine if any unauthorized access has occurred. For each drawing, the Lotto Texas balls are mixed by four acrylic mixing paddles rotating clockwise. High speed is used for mixing and low speed for ball selection. As each ball is selected, it rolls down a chute into an official number display area. We wish to compute the probability of winning the Mega Millions Grand Prize, which requires the correct selection of the five white balls plus the gold Mega ball.*

The probability of winning the Mega Millions Grand Prize is

$$\frac{1}{\binom{56}{5}}\frac{1}{46} = \frac{51!5!}{56!}\frac{1}{46} = \frac{1}{175711536}.$$

Example 15 (Sinking Boat). *Six couples, twelve people total, are out at sea on a sail boat. Unfortunately, the boat hits an iceberg and starts sinking slowly. Two Coast Guard vessels, the Ibis and the Mako, come to the rescue. Each boat can rescue six people. What is the probability that no two people from a same couple end up on the Mako?*

Suppose that rescued passengers are assigned to the Ibis and the Mako at random. Then, the number of possible ways to partition these passengers between the two vessels is

$$\binom{12}{6} = \frac{12!}{6!6!}.$$

If no two people from a same couple end up on the Mako, then each couple is split between the two vessels. In these circumstances, there are two possibilities for every couple and, using the counting principle, we gather that there are 2^6 *such assignments. Collecting these results, we conclude that the probability that no two people from a same couple end up on the Mako is equal to*

$$\frac{2^6}{\binom{12}{6}} = \frac{2^6 6!6!}{12!}.$$

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Chapter 1.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 1.6.

3. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Section 1.7.

Chapter 3

Basic Concepts of Probability

The theory of probability provides the mathematical tools necessary to study and analyze uncertain phenomena that occur in nature. It establishes a formal framework to understand and predict the outcome of a random experiment. It can be used to model complex systems and characterize stochastic processes. This is instrumental in designing efficient solutions to many engineering problems. Two components define a probabilistic model, a sample space and a probability law.

3.1 Sample Spaces and Events

In the context of probability, an *experiment* is a random occurrence that produces one of several *outcomes*. The set of all possible outcomes is called the *sample space* of the experiment, and it is denoted by Ω. An *admissible* subset of the sample space is called an *event*.

Example 16. *The rolling of a die forms a common experiment. A sample space* Ω *corresponding to this experiment is given by the six faces of a die. The set of prime numbers less than or equal to six, namely* $\{2, 3, 5\}$*, is one of many possible events. The actual number observed after rolling the die is the outcome of the experiment.*

There is essentially no restriction on what constitutes an experiment. The flipping of a coin, the flipping of n coins, and the tossing of an infinite sequence of coins are all random experiments. Also, two similar experiments may have

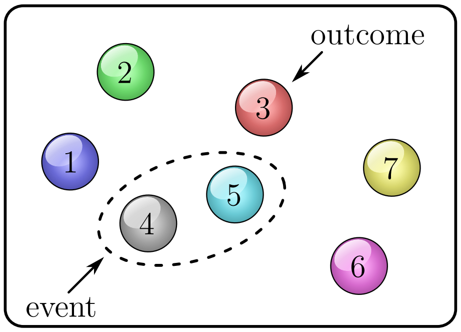

Figure 3.1: A sample space contains all the possible outcomes; an admissible subset of the sample space is called an event.

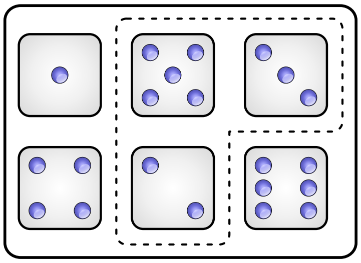

Figure 3.2: A possible sample space for the rolling of a die is $\Omega = \{1, 2, \ldots, 6\}$, and the subset $\{2, 3, 5\}$ forms a specific event.

different sample spaces. A sample space Ω for observing the number of heads in n tosses of a coin is $\{0, 1, \ldots, n\}$; however, when describing the complete history of the n coin tosses, the sample space becomes much larger with 2^n distinct sequences of heads and tails. Ultimately, the choice of a particular sample space depends on the properties one wishes to analyze. Yet some rules must be followed in selecting a sample space.

1. The elements of a sample space should be *distinct* and *mutually exclusive*. This ensures that the outcome of an experiment is unique.

2. A sample space must be *collectively exhaustive*. That is, every possible outcome of the experiment must be accounted for in the sample space.

In general, a sample space should be precise enough to distinguish between all outcomes of interest, while avoiding frivolous details.

Example 17. *Consider the space composed of the odd integers located between one and six, the even integers contained between one and six, and the prime numbers less than or equal to six. This collection cannot be a sample space for the rolling of a die because its elements are not mutually exclusive. In particular, the numbers three and five are both odd and prime, while the number two is prime and even. This violates the uniqueness criterion.*

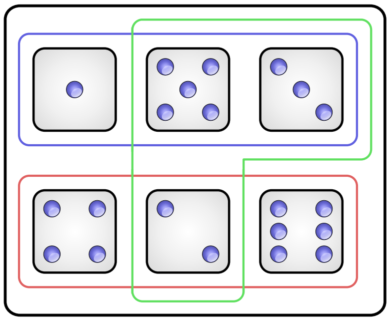

Figure 3.3: This collection of objects cannot be a sample space as the three proposed outcomes (even, odd and prime) are not mutually exclusive.

Alternatively, the elements of the space composed of the odd numbers between one and six, and the even numbers between one and six, are distinct and mutually exclusive; an integer cannot be simultaneously odd and even. Moreover, this space is collectively exhaustive because every integer is either odd or even. This latter description forms a possible sample space for the rolling of a die.

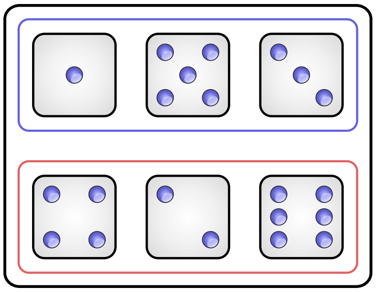

Figure 3.4: A candidate sample space for the rolling of a die is composed of two objects, the odd numbers and the even numbers between one and six.

3.2 Probability Laws

A *probability law* specifies the likelihood of events related to an experiment. Formally, a probability law assigns to every event A a number $\Pr(A)$, called the *probability of event* A, such that the following axioms are satisfied.

1. **(Nonnegativity)** $\Pr(A) \geq 0$, for every event A.
2. **(Normalization)** The probability of the sample space Ω is equal to one,
$$\Pr(\Omega) = 1.$$
3. **(Countable Additivity)** If A and B are disjoint events, $A \cap B = \emptyset$, then the probability of their union satisfies
$$\Pr(A \cup B) = \Pr(A) + \Pr(B).$$
More generally, if $A_1, A_2, \ldots$ is a sequence of disjoint events and $\bigcup_{k=1}^{\infty} A_k$ is itself an admissible event then
$$\Pr\left(\bigcup_{k=1}^{\infty} A_k \right) = \sum_{k=1}^{\infty} \Pr(A_k).$$

A number of important properties can be deduced from the three axioms of probability. We prove two such properties below. The first statement describes the relation between inclusion and probabilities.

Proposition 1. *If* $A \subset B$, *then* $\Pr(A) \leq \Pr(B)$.

Proof. Since $A \subset B$, we have $B = A \cup (B - A)$. Noting that A and $B - A$ are disjoint sets, we get
$$\Pr(B) = \Pr(A) + \Pr(B - A) \geq \Pr(A),$$
where the inequality follows from the nonnegativity of probability laws. □

Our second result specifies the probability of the union of two events that are not necessarily mutually exclusive.

Proposition 2. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

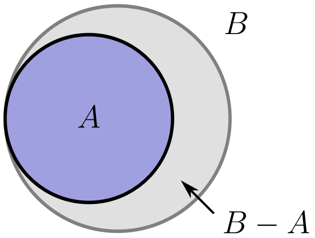

Figure 3.5: If event A is a subset of event B, then the probability of A is less than or equal to the probability of B.

Proof. Using the third axiom of probability on the disjoint sets A and $(A \cup B) - A$, we can write

$$\Pr(A \cup B) = \Pr(A) + \Pr((A \cup B) - A) = \Pr(A) + \Pr(B - A).$$

Similarly, applying the third axiom to $A \cap B$ and $B - (A \cap B)$, we obtain

$$\Pr(B) = \Pr(A \cap B) + \Pr(B - (A \cap B)) = \Pr(A \cap B) + \Pr(B - A).$$

Combining these two equations yields the desired result. □

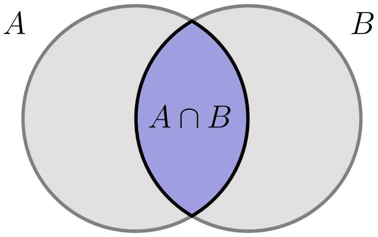

Figure 3.6: The probability of the union of A and B is equal to the probability of A plus the probability of B minus the probability of their intersection.

The statement of Proposition 2 can be extended to finite unions of events. Specifically, we can write

$$\Pr\left(\bigcup_{k=1}^{n} A_k \right) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{\mathbb{I} \subset \{1,\ldots,n\}, |\mathbb{I}|=k} \Pr\left(\bigcap_{i \in \mathbb{I}} A_i \right)$$

where the rightmost sum runs over all subsets $\mathbb{I}$ of $\{1, \ldots, n\}$ that contain exactly k elements. This more encompassing result is known as the *inclusion-exclusion principle*.

We can use Proposition 2 recursively to derive a bound on the probabilities of unions. This theorem, which is sometimes called the *Boole inequality*, asserts that the probability of at least one event occurring is no greater than the sum of the probabilities of the individual events.

Theorem 1 (Union Bound)**.** *Let $A_1, A_2, \ldots, A_n$ be a collection of events, then*

$$\Pr\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \Pr(A_k). \tag{3.1}$$

Proof. We show this result using induction. First, we note that the claim is trivially true for $n = 1$. As an inductive hypothesis, assume that (3.1) holds for some $n \geq 1$. Then, we have

$$\begin{aligned}
\Pr\left(\bigcup_{k=1}^{n+1} A_k\right) &= \Pr\left(A_{n+1} \cup \left(\bigcup_{k=1}^{n} A_k\right)\right) \\
&= \Pr(A_{n+1}) + \Pr\left(\bigcup_{k=1}^{n} A_k\right) - \Pr\left(A_{n+1} \cap \left(\bigcup_{k=1}^{n} A_k\right)\right) \\
&\leq \Pr(A_{n+1}) + \Pr\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n+1} \Pr(A_k).
\end{aligned}$$

Therefore, by the principle of mathematical induction, (3.1) is valid for all positive integers. □

The union bound is often employed in situations where finding the joint probability of multiple rare events is difficult, but computing the probabilities of the individual components is straightforward.

Example 18. *An urn contains 990 blue balls and 10 red balls. Five people each pick a ball at random, without replacement. We wish to compute the probability that at least one person picks a red ball. Let B_k denote the event that person k draws a red ball. We note that the probability of interest can be written as $\Pr\left(\bigcup_{k=1}^{5} B_k\right)$.*

We first approximate this probability using the union bound. The probability that a particular person picks a red ball is equal to 1/100. Applying (3.1)*, we get a bound on the probability that at least one person picks a red ball,*

$$\Pr\left(\bigcup_{k=1}^{5} B_k\right) \leq \sum_{k=1}^{5} \Pr(B_k) = \sum_{k=1}^{5} \frac{1}{100} = \frac{1}{20}.$$

We can also compute this probability exactly. The probability that no red balls are selected is given by $\binom{990}{5}/\binom{1000}{5}$. *Hence, the probability that a least one person draws a red ball becomes*

$$\Pr\left(\bigcup_{k=1}^{5} B_k\right) = 1 - \frac{\binom{990}{5}}{\binom{1000}{5}} \approx 0.0491.$$

As a second application, consider the problem where the five people each draw two balls from the urn, without replacement. This time, we wish to approximate the probability that at least one person gets two red balls. Using the same steps as before, we get

$$\Pr\left(\bigcup_{k=1}^{5} C_k\right) \leq \sum_{k=1}^{5} \Pr(C_k) = \sum_{k=1}^{5} \frac{\binom{10}{2}}{\binom{1000}{2}} = \frac{1}{2220},$$

where C_k *represents the event that person* k *draws two red balls. In this latter scenario, computing the exact probability is much more challenging.*

3.2.1 Finite Sample Spaces

If a sample space Ω contains a finite number of elements, then a probability law on Ω is completely determined by the probabilities of its individual outcomes. Denote a sample space containing n elements by $\Omega = \{s_1, s_2, \ldots, s_n\}$. Any event in Ω is of the form $A = \{s_i \in \Omega | i \in \mathbb{I}\}$, where $\mathbb{I}$ is a subset of the integers one through n. The probability of event A is therefore given by the third axiom of probability,

$$\Pr(A) = \Pr(\{s_i \in \Omega | i \in \mathbb{I}\}) = \sum_{i \in \mathbb{I}} \Pr(s_i).$$

We emphasize that, by definition, distinct outcomes are always disjoint events.

If in addition the elements of Ω are equally likely with

$$\Pr(s_1) = \Pr(s_2) = \cdots = \Pr(s_n) = \frac{1}{n},$$

then the probability of an event A becomes

$$\Pr(A) = \frac{|A|}{n} \tag{3.2}$$

where $|A|$ denotes the number of elements in A.

Example 19. *The rolling of a fair die is an experiment with a finite number of equally likely outcomes. The probability of observing any of the faces labeled one through six is therefore equal to* $1/6$*. The probability of any event can easily be computed by counting the number of distinct outcomes included in the event. For instance, the probability of rolling a prime number is*

$$\Pr(\{2,3,5\}) = \Pr(2) + \Pr(3) + \Pr(5) = \frac{3}{6}.$$

3.2.2 Countably Infinite Models

Consider a sample space that consists of a countably infinite number of elements, $\Omega = \{s_1, s_2, \ldots\}$. Again, a probability law on Ω is specified by the probabilities of individual outcomes. An event in Ω can be written as $A = \{s_j \in \Omega | j \in \mathbb{J}\}$, where $\mathbb{J}$ is a subset of the positive integers. Using the third axiom of probability, $\Pr(A)$ can be written as

$$\Pr(A) = \Pr(\{s_j \in \Omega | j \in \mathbb{J}\}) = \sum_{j \in \mathbb{J}} \Pr(s_j).$$

The possibly infinite sum $\sum_{j \in \mathbb{J}} \Pr(s_j)$ always converges since the summands are nonnegative and the sum is bounded above by one; it is consequently well defined.

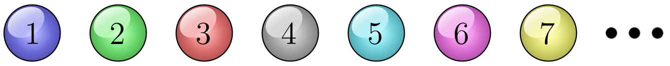

Figure 3.7: A countable set is a collection of elements with the same cardinality as some subset of the natural numbers.

Example 20. *Suppose that a fair coin is tossed repetitively until heads is observed. The number of coin tosses is recorded as the outcome of this experiment. A natural sample space for this experiment is* $\Omega = \{1, 2, \ldots\}$*, a countably infinite set.*

The probability of observing heads on the first trial is $1/2$*, and the probability of observing heads for the first time on trial* k *is* 2^{-k}*. The probability of the entire sample space is therefore equal to*

$$\Pr(\Omega) = \sum_{k=1}^{\infty} \Pr(k) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1,$$

as expected. Similarly, the probability of the number of coin tosses being even can be computed as

$$\Pr(\{2,4,6,\ldots\}) = \sum_{k=1}^{\infty} \Pr(2k) = \sum_{k=1}^{\infty} \frac{1}{2^{2k}} = \frac{1}{4} \frac{1}{\left(1-\frac{1}{4}\right)} = \frac{1}{3}.$$

3.2.3 Uncountably Infinite Models

Probabilistic models with uncountably infinite sample spaces differ from the finite and countable cases in that a probability law may not necessarily be specified by the probabilities of single-element outcomes. This difficulty arises from the large number of elements contained in the sample space when the latter is uncountable. Many subsets of Ω do not have a finite or countable representation, and as such the third axiom of probability cannot be applied to relate the probabilities of these events to the probabilities of individual outcomes. Despite these apparent difficulties, probabilistic models with uncountably infinite sample spaces are quite useful in practice. To develop an understanding of uncountable probabilistic models, we consider the unit interval $[0,1]$.

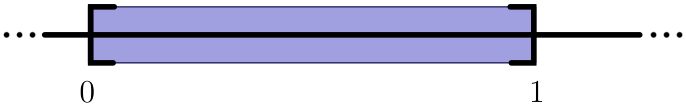

Figure 3.8: The unit interval $[0,1]$, composed of all real numbers between zero and one, is an example of an uncountable set.

Suppose that an element is chosen at random from this interval, with uniform weighting. By the first axiom of probability, the probability that this element belongs to the interval $[0,1]$ is given by $\Pr([0,1]) = 1$. Furthermore, if two intervals have the same length, the probabilities of the outcome falling in either interval should be identical. For instance, it is natural to anticipate that $\Pr((0,0.25)) = \Pr((0.75,1))$.

In an extension of the previous observation, we take the probability of an open interval (a,b) where $0 \leq a < b \leq 1$ to equal

$$\Pr((a,b)) = b - a. \tag{3.3}$$

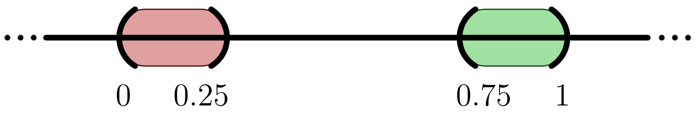

Figure 3.9: If the outcome of an experiment is uniformly distributed over $[0, 1]$, then two subintervals of equal lengths should have the same probabilities.

Using the third axiom of probability, it is then possible to find the probability of a finite or countable union of disjoint open intervals.

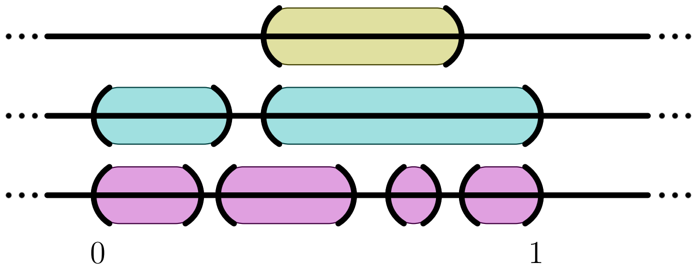

Figure 3.10: The probabilities of events that are formed through the union of disjoint intervals can be computed in a straightforward manner.

Specifically, for constants $0 \leq a_1 < b_1 < a_2 < b_2 < \cdots \leq 1$, we get

$$\Pr\left(\bigcup_{k=1}^{\infty}(a_k, b_k)\right) = \sum_{k=1}^{\infty}(b_k - a_k).$$

The probabilities of more complex events can be obtained by applying additional elementary set operations. However, it suffices to say for now that specifying the probability of the outcome falling in (a, b) for every possible open interval is enough to define a probability law on Ω. In the example at hand, (3.3) completely determines the probability law on $[0, 1]$.

Note that we can give an alternative means of computing the probability of an interval. Again, consider the open interval (a, b) where $0 \leq a < b \leq 1$. The probability of the outcome falling in this interval is equal to

$$\Pr((a, b)) = b - a = \int_a^b dx = \int_{(a,b)} dx.$$

Moreover, for $0 \leq a_1 < b_1 < a_2 < b_2 < \cdots \leq 1$, we can write

$$\Pr\left(\bigcup_{k=1}^{\infty} (a_k, b_k) \right) = \sum_{k=1}^{\infty} (b_k - a_k) = \int_{\bigcup_{k=1}^{\infty} (a_k, b_k)} dx.$$

For this carefully crafted example, it appears that the probability of an admissible event A is given by the integral

$$\Pr(A) = \int_A dx.$$

This is indeed accurate for the current scenario. In fact, the class of admissible events for this experiment is simply the collection of all sets for which the integral $\int_A dx$ can be computed. In other words, if a number is chosen at random from $[0, 1]$, then the probability of this number falling in set $A \subset [0, 1]$ is

$$\Pr(A) = \int_A dx.$$

This method of computing probabilities can be extended to more complicated problems. In these notes, we will see many probabilistic models with uncountably infinite sample spaces. The mathematical tools required to handle such models will be treated alongside.

Example 21. *Suppose that a participant at a game-show is required to spin the wheel of serendipity, a perfect circle with unit radius. When subjected to a vigorous spin, the wheel is equally likely to stop anywhere along its perimeter. A sampling space for this experiment is the collection of all angles from* 0 *to* 2π*, an uncountable set. The probability of* Ω *is invariably equal to one,* $\Pr([0, 2\pi)) = 1$.

The probability that the wheel stops in the first quadrant is given by

$$\Pr\left(\left[0, \frac{\pi}{2} \right) \right) = \int_0^{\frac{\pi}{2}} \frac{1}{2\pi} d\theta = \frac{1}{4}.$$

More generally, the probability that the wheel stops in an interval (a, b) *where* $0 \leq a \leq b < 2\pi$ *can be written as*

$$\Pr((a, b)) = \frac{b - a}{2\pi}.$$

If $B \subset [0, 2\pi)$ *is a set representing all winning outcomes, then the probability of success at the wheel becomes*

$$\Pr(B) = \int_B \frac{1}{2\pi} d\theta.$$

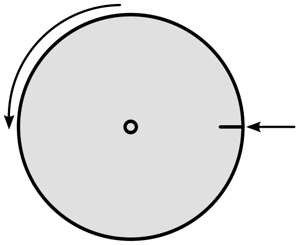

Figure 3.11: The wheel of serendipity forms an example of a random experiment for which the sample space is uncountable.

3.2.4 Probability and Measure Theory*

A thorough treatment of probability involves advanced mathematical concepts, especially when it comes to infinite sample spaces. The basis of our intuition for the infinite is the set of *natural numbers*,

$$\mathbb{N} = \{1, 2, \ldots\}.$$

Two sets are said to have the same *cardinality* if their elements can be put in one-to-one correspondence. A set with the same cardinality as a subset of the natural numbers is said to be *countable*. That is, the elements of a countable set can always be listed in sequence, $s_1, s_2, \ldots$; although the order may have nothing to do with any relation between the elements. The integers and the rational numbers are examples of countably infinite sets. It may be surprising at first to learn that there exist uncountable sets. To escape beyond the countable, one needs set theoretic tools such as *power sets*. The set of real numbers is uncountably infinite; it cannot be put in one-to-one correspondence with the natural numbers. A typical progression in analysis consists of using the finite to gain intuition about the countably infinite, and then to employ the countably infinite to get at the uncountable.

It is tempting to try to assign probabilities to every subset of a sample space Ω. However, for uncountably infinite sample spaces, this leads to serious difficulties that cannot be resolved. In general, it is necessary to work with special subclasses of the class of all subsets of a sample space Ω. The collections of the appropriate kinds are called fields and σ-fields, and they are studied in *measure theory*. This leads to measure-theoretic probability, and to its unified treatment of the discrete and the continuous.

Fortunately, it is possible to develop a working understanding of probability without worrying excessively about these issues. At some point in your academic career, you may wish to study analysis and measure theory more carefully and in greater details. However, it is not our current purpose to initiate the rigorous treatment of these topics.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Chapter 2.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 1.2.

3. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Sections 2.1–2.3.

4. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 1.1,1.3–1.4.

Chapter 4

Conditional Probability

Conditional probability provides a way to compute the likelihood of an event based on *partial information*. This is a powerful concept that is used extensively throughout engineering with applications to decision making, networks, communications and many other fields.

4.1 Conditioning on Events

We begin our description of conditional probability with illustrative examples. The intuition gained through this exercise is then generalized by introducing a formal definition for this important concept.

Example 22. *The rolling of a fair die is an experiment with six equally likely outcomes. As such, the probability of obtaining any of the outcomes is* $1/6$*. However, if we are told that the upper face features an odd number, then only three possibilities remain, namely* $\{1, 3, 5\}$*. These three outcomes had equal probabilities before the additional information was revealed. It then seems natural to assume that they remain equally likely afterwards. In particular, it is reasonable to assign a probability of* $1/3$ *to each of the three outcomes that remain possible candidates after receiving the side information. We can express the probability of getting a three given that the outcome is an odd number as*

$$\frac{\Pr(3 \cap \{1, 3, 5\})}{\Pr(\{1, 3, 5\})} = \frac{\Pr(3)}{\Pr(\{1, 3, 5\})} = \frac{1}{3}.$$

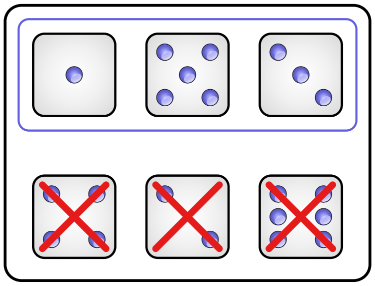

Figure 4.1: Partial information about the outcome of an experiment may change the likelihood of events. The resulting values are known as conditional probabilities.

Example 23. *Let A and B be events associated with a random experiment, and assume that* $\Pr(B) > 0$. *To gain additional insight into conditional probability, we consider the scenario where this experiment is repeated N times. Let N_{AB} be the number of trials for which $A \cap B$ occurs, $N_{A\overline{B}}$ be the number of times where only A occurs, $N_{\overline{A}B}$ be the number of times where only B occurs, and $N_{\overline{AB}}$ be the number of trials for which neither takes place. From these definitions, we gather that A is observed exactly $N_A = N_{AB} + N_{A\overline{B}}$ times, and B is seen $N_B = N_{AB} + N_{\overline{A}B}$ times.*

The frequentist view of probability is based on the fact that, as N becomes large, one can approximate the probability of an event by taking the ratio of the number of times this event occurs over the total number of trials. For instance, we can write

$$\Pr(A \cap B) \approx \frac{N_{AB}}{N} \qquad\qquad \Pr(B) \approx \frac{N_B}{N}.$$

Likewise, the conditional probability of A given knowledge that the outcome lies in B can be computed using

$$\Pr(A|B) \approx \frac{N_{AB}}{N_B} = \frac{N_{AB}/N}{N_B/N} \approx \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{4.1}$$

As N approaches infinity, these approximations become exact and (4.1) *unveils the formula for conditional probability.*

Having considered intuitive arguments, we turn to the mathematical definition of conditional probability. Let B be an event such that $\Pr(B) > 0$.

A conditional probability law assigns to every event A a number $\Pr(A|B)$, termed the *conditional probability of A given B*, such that

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{4.2}$$

We can show that the collection of conditional probabilities $\{\Pr(A|B)\}$ specifies a valid probability law, as defined in Section 3.2. For every event A, we have

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \geq 0$$

and, hence, $\Pr(A|B)$ is nonnegative. The probability of the entire sample space Ω is equal to

$$\Pr(\Omega|B) = \frac{\Pr(\Omega \cap B)}{\Pr(B)} = \frac{\Pr(B)}{\Pr(B)} = 1.$$

If $A_1, A_2, \ldots$ is a sequence of disjoint events, then

$$A_1 \cap B, A_2 \cap B, \ldots$$

is also a sequence of disjoint events and

$$\begin{aligned}\Pr\left(\bigcup_{k=1}^{\infty} A_k \middle| B \right) &= \frac{\Pr\left(\left(\bigcup_{k=1}^{\infty} A_k \right) \cap B \right)}{\Pr(B)} = \frac{\Pr\left(\bigcup_{k=1}^{\infty} (A_k \cap B) \right)}{\Pr(B)} \\ &= \sum_{k=1}^{\infty} \frac{\Pr(A_k \cap B)}{\Pr(B)} = \sum_{k=1}^{\infty} \Pr(A_k|B),\end{aligned}$$

where the third equality follows from the third axiom of probability applied to the set $\bigcup_{k=1}^{\infty} (A_k \cap B)$. Thus, the conditional probability law defined by (4.2) satisfies the three axioms of probability.

Example 24. *A fair coin is tossed repetitively until heads is observed. In Example 20, we found that the probability of observing heads for the first time on trial k is 2^{-k}. We now wish to compute the probability that heads occurred for the first time on the second trial given that it took an even number of tosses to observe heads. In this example, $A = \{2\}$ and B is the set of even numbers. The probability that the outcome is two, given that the number of tosses is even, is equal to*

$$\Pr(2|B) = \frac{\Pr(2 \cap B)}{\Pr(B)} = \frac{\Pr(2)}{\Pr(B)} = \frac{1/4}{1/3} = \frac{3}{4}.$$

In the above computation, we have used the fact that the probability of flipping the coin an even number of times is equal to $1/3$*. This fact was established in Example 20.*

The definition of conditional probability can be employed to compute the probability of several events occurring simultaneously. Let $A_1, A_2, \ldots, A_n$ be a collection of events. The probability of events A_1 through A_n taking place at the same time is given by

$$\Pr\left(\bigcap_{k=1}^{n} A_k \right) = \Pr(A_1) \Pr(A_2 | A_1) \Pr(A_3 | A_1 \cap A_2) \cdots \Pr\left(A_n \middle| \bigcap_{k=1}^{n-1} A_k \right). \quad (4.3)$$

This formula is known as the *chain rule of probability*, and it can be verified by expanding each of the conditional probabilities using (4.2),

$$\Pr\left(\bigcap_{k=1}^{n} A_k \right) = \Pr(A_1) \frac{\Pr(A_1 \cap A_2)}{\Pr(A_1)} \frac{\Pr(A_1 \cap A_2 \cap A_3)}{\Pr(A_1 \cap A_2)} \cdots \frac{\Pr\left(\bigcap_{k=1}^{n} A_k \right)}{\Pr\left(\bigcap_{k=1}^{n-1} A_k \right)}.$$

This latter expression implicitly assumes that $\Pr\left(\bigcap_{k=1}^{n-1} A_k \right) \neq 0$.

Example 25. *An urn contains eight blue balls and four green balls. Three balls are drawn from this urn without replacement. We wish to compute the probability that all three balls are blue. The probability of drawing a blue ball the first time is equal to* $8/12$*. The probability of drawing a blue ball the second time given that the first ball is blue is* $7/11$*. Finally, the probability of drawing a blue ball the third time given that the first two balls are blue is* $6/10$*. Using* (4.3)*, we can compute the probability of drawing three blue balls as*

$$\Pr(bbb) = \frac{8}{12} \frac{7}{11} \frac{6}{10} = \frac{14}{55}.$$

4.2 The Total Probability Theorem

The probability of events A and B occurring at the same time can be calculated as a special case of (4.3). For two events, this computational formula simplifies to

$$\Pr(A \cap B) = \Pr(A|B) \Pr(B). \quad (4.4)$$

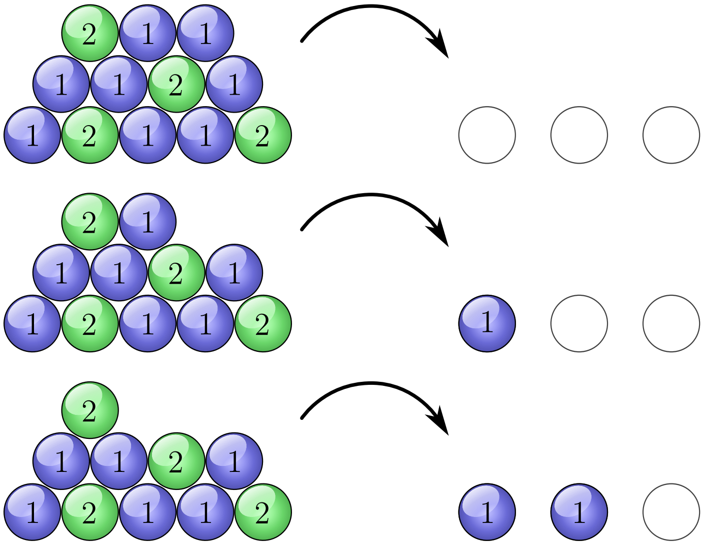

Figure 4.2: Conditional probability can be employed to calculate the probability of multiple events occurring at the same time.

We can also obtain this equation directly from the definition of conditional probability. This property is a key observation that plays a central role in establishing two important results, the *total probability theorem* and *Bayes' rule*. To formulate these two theorems, we need to revisit the notion of a partition. A collection of events $A_1, A_2, \ldots, A_n$ is said to be a *partition* of the sample space Ω if these events are disjoint and their union is the entire sample space,

$$\bigcup_{k=1}^{n} A_k = \Omega.$$

Visually, a partition divides an entire set into disjoint subsets, as exemplified in Figure 4.3.

Theorem 2 (Total Probability Theorem)**.** *Let $A_1, A_2, \ldots, A_n$ be a collection of events that forms a partition of the sample space Ω. Suppose that $\Pr(A_k) > 0$ for all k. Then, for any event B, we can write*

$$\begin{aligned}\Pr(B) &= \Pr(A_1 \cap B) + \Pr(A_2 \cap B) + \cdots + \Pr(A_n \cap B) \\ &= \Pr(A_1) \Pr(B|A_1) + \Pr(A_2) \Pr(B|A_2) + \cdots + \Pr(A_n) \Pr(B|A_n).\end{aligned}$$

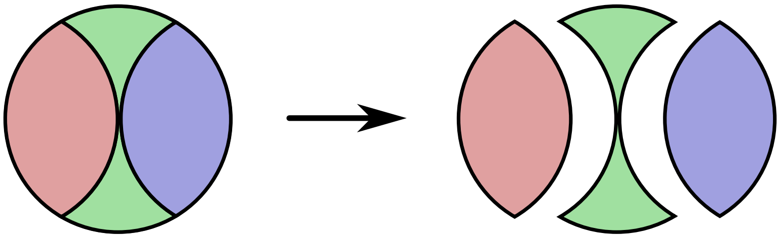

Figure 4.3: A partition of S can be formed by selecting a collection of subsets that are disjoint and whose union is S.

Proof. The collection of events $A_1, A_2, \ldots, A_n$ forms a partition of the sample space Ω. We can therefore write

$$B = B \cap \Omega = B \cap \left(\bigcup_{k=1}^{n} A_k \right).$$

Since $A_1, A_2, \ldots, A_n$ are disjoint sets, the events $A_1 \cap B, A_2 \cap B, \ldots, A_n \cap B$ are also disjoint. Combining these two facts, we get

$$\begin{aligned} \Pr(B) &= \Pr\left(B \cap \left(\bigcup_{k=1}^{n} A_k \right) \right) = \Pr\left(\bigcup_{k=1}^{n} (B \cap A_k) \right) \\ &= \sum_{k=1}^{n} \Pr\left(B \cap A_k \right) = \sum_{k=1}^{n} \Pr(A_k) \Pr\left(B | A_k \right), \end{aligned}$$

where the fourth equality follows from the third axiom of probability. □

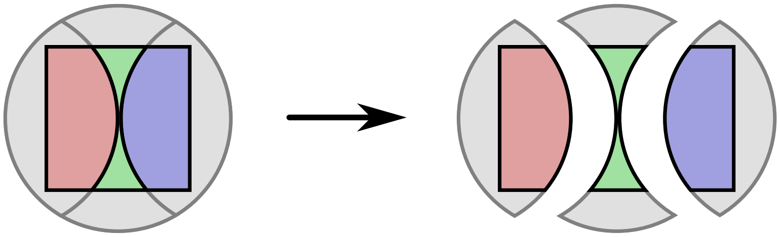

Figure 4.4: The total probability theorem states that the probability of event B can be computed by summing $\Pr(A_i \cap B)$ over all members of the partition $A_1, A_2, \ldots, A_n$.

A graphical interpretation of Theorem 2 is illustrated in Figure 4.4. Event B can be decomposed into the disjoint union of $A_1 \cap B, A_2 \cap B, \ldots, A_n \cap B$.

The probability of event B can then be computed by adding the corresponding summands

$$\Pr(A_1 \cap B), \Pr(A_2 \cap B), \ldots, \Pr(A_n \cap B).$$

Example 26. *An urn contains five green balls and three red balls. A second urn contains three green balls and nine red balls. One of the two urns is picked at random, with equal probabilities, and a ball is drawn from the selected urn. We wish to compute the probability of obtaining a green ball.*

In this problem, using a divide and conquer approach seems appropriate; we therefore utilize the total probability theorem. If the first urn is chosen, then the ensuing probability of getting a green ball is $5/8$*. One the other hand, if a ball is drawn from the second urn, the probability that it is green reduces to* $3/12$*. Since the probability of selecting either urn is* $1/2$*, we can write the overall probability of getting a green ball as*

$$\begin{aligned}\Pr(g) &= \Pr(g \cap U_1) + \Pr(g \cap U_2) \\ &= \Pr(g|U_1)\Pr(U_1) + \Pr(g|U_2)\Pr(U_2) \\ &= \frac{5}{8} \cdot \frac{1}{2} + \frac{3}{12} \cdot \frac{1}{2} = \frac{7}{16}.\end{aligned}$$

4.3 Bayes' Rule

The following result is also very useful. It relates the conditional probability of A given B to the conditional probability of B given A.

Theorem 3 (Bayes' Rule)**.** *Let* $A_1, A_2, \ldots, A_n$ *be a collection of events that forms a partition of the sample space* Ω*. Suppose that* $\Pr(A_k) > 0$ *for all* k*. Then, for any event* B *such that* $\Pr(B) > 0$*, we can write*

$$\begin{aligned}\Pr(A_i|B) &= \frac{\Pr(A_i)\Pr(B|A_i)}{\Pr(B)} \\ &= \frac{\Pr(A_i)\Pr(B|A_i)}{\sum_{k=1}^{n} \Pr(A_k)\Pr(B|A_k)}.\end{aligned} \tag{4.5}$$

Proof. Bayes' rule is easily verified. We expand the probability of $A_i \cap B$ using (4.4) twice, and we get

$$\Pr(A_i \cap B) = \Pr(A_i|B)\Pr(B) = \Pr(B|A_i)\Pr(A_i).$$

Rearranging the terms yields the first equality. The second equality in (4.5) is obtained by applying Theorem 2 to the denominator $\Pr(B)$. □

Example 27. *April, a biochemist, designs a test for a latent disease. If a subject has the disease, the probability that the test results turn out positive is* 0.95*. Similarly, if a subject does not have the disease, the probability that the test results come up negative is* 0.95*. Suppose that one percent of the population is infected by the disease. We wish to find the probability that a person who tested positive has the disease.*

Let D denote the event that a person has the disease, and let P be the event that the test results are positive. Using Bayes' rule, we can compute the probability that a person who tested positive has the disease,

$$\begin{aligned}\Pr(D|P) &= \frac{\Pr(D)\Pr(P|D)}{\Pr(D)\Pr(P|D) + \Pr(D^{\mathrm{c}})\Pr(P|D^{\mathrm{c}})} \\ &= \frac{0.01 \cdot 0.95}{0.01 \cdot 0.95 + 0.99 \cdot 0.05} \\ &\approx 0.1610.\end{aligned}$$

Although the test may initially appear fairly accurate, the probability that a person with a positive test carries the disease remains small.

4.4 Independence

Two events A and B are said to be *independent* if $\Pr(A \cap B) = \Pr(A)\Pr(B)$. Interestingly, independence is closely linked to the concept of conditional probability. If $\Pr(B) > 0$ and events A and B are independent, then

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

That is, the *a priori* probability of event A is identical to the *a posteriori* probability of A given B. In other words, if A is independent of B, then partial knowledge of B contains no information about the likely occurrence of A. We note that independence is a symmetric relation; if A is independent of B, then B is also independent of A. It is therefore unambiguous to say that A and B are independent events.

Example 28. *Suppose that two dice are rolled at the same time, a red die and a blue die. We observe the numbers that appear on the upper faces of the two dice. The sample space for this experiment is composed of thirty-six equally likely outcomes. Consider the probability of getting a four on the red die given that the blue die shows a six,*

$$\begin{aligned}\Pr(\{r=4\}|\{b=6\}) &= \frac{\Pr(\{r=4\}\cap\{b=6\})}{\Pr(b=6)}\\ &= \frac{1}{6} = \Pr(r=4).\end{aligned}$$

From this equation, we gather that

$$\Pr(\{r=4\}\cap\{b=6\}) = \Pr(r=4)\Pr(b=6).$$

As such, rolling a four on the red die and rolling a six on the blue die are independent events.

Similarly, consider the probability of obtaining a four on the red die given that the sum of the two dice is eleven,

$$\begin{aligned}\Pr(\{r=4\}|\{r+b=11\}) &= \frac{\Pr(\{r=4\}\cap\{r+b=11\})}{\Pr(r+b=11)} = 0\\ &\neq \frac{1}{6} = \Pr(r=4).\end{aligned}$$

In this case, we conclude that getting a four on the red die and a sum total of eleven are not independent events.

The basic idea of independence seems intuitively clear: if knowledge about the occurrence of event B has no impact on the probability of A, then these two events must be independent. Yet, independent events are not necessarily easy to visualize in terms of their sample space. A common mistake is to assume that two events are independent if they are disjoint. Two mutually exclusive events can hardly be independent: if $\Pr(A) > 0$, $\Pr(B) > 0$, and $\Pr(A \cap B) = 0$ then

$$\Pr(A \cap B) = 0 < \Pr(A)\Pr(B).$$

Hence, A and B cannot be independent if they are disjoint, non-trivial events.

4.4.1 Independence of Multiple Events

The concept of independence can be extended to multiple events. The events $A_1, A_2, \ldots, A_n$ are *independent* provided that

$$\Pr \left(\bigcap_{i \in \mathbb{I}} A_i \right) = \prod_{i \in \mathbb{I}} \Pr(A_i), \tag{4.6}$$

for every subset $\mathbb{I}$ of $\{1, 2, \ldots, n\}$.

For instance, consider a collection of three events, A, B and C. These events are independent whenever

$$\begin{aligned} \Pr(A \cap B) &= \Pr(A) \Pr(B) \\ \Pr(A \cap C) &= \Pr(A) \Pr(C) \\ \Pr(B \cap C) &= \Pr(B) \Pr(C) \end{aligned} \tag{4.7}$$

and, in addition,

$$\Pr(A \cap B \cap C) = \Pr(A) \Pr(B) \Pr(C).$$

The three equalities in (4.7) assert that A, B and C are *pairwise independent*. Note that the fourth equation does not follow from the first three conditions, nor does it imply any of them. Pairwise independence does not necessarily imply independence. This is illustrated below.

Example 29. *A fair coin is flipped twice. Let A denote the event that heads is observed on the first toss. Let B be the event that heads is obtained on the second toss. Finally, let C be the event that the two coins show distinct sides. These three events each have a probability of $1/2$. Furthermore, we have*

$$\Pr(A \cap B) = \Pr(A \cap C) = \Pr(B \cap C) = \frac{1}{4}$$

and, therefore, these events are pairwise independent. However, we can verify that

$$\Pr(A \cap B \cap C) = 0 \neq \frac{1}{8} = \Pr(A) \Pr(B) \Pr(C).$$

This shows that events A, B and C are not independent.

Example 30. *Two dice are rolled at the same time, a red die and a blue die. Let A be the event that the number on the red die is odd. Let B be the event that the number on the red die is either two, three or four. Also, let C be the event that the product of the two dice is twelve. The individual probabilities of these events are*

$$\begin{aligned}\Pr(A) &= \Pr(r \in \{1,3,5\}) = \frac{1}{2}\\ \Pr(B) &= \Pr(r \in \{2,3,4\}) = \frac{1}{2}\\ \Pr(C) &= \Pr(r \times b = 12) = \frac{4}{36}.\end{aligned}$$

We note that these events are not pairwise independent because

$$\begin{aligned}\Pr(A \cap B) &= \frac{1}{6} \neq \frac{1}{4} = \Pr(A)\Pr(B)\\ \Pr(A \cap C) &= \frac{1}{36} \neq \frac{1}{18} = \Pr(A)\Pr(C)\\ \Pr(B \cap C) &= \frac{1}{12} \neq \frac{1}{18} = \Pr(B)\Pr(C).\end{aligned}$$

Consequently, the multiple events A, B and C are not independent. Still, the probability of these three events occurring simultaneously is

$$\Pr(A \cap B \cap C) = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{36} = \Pr(A)\Pr(B)\Pr(C).$$

4.4.2 Conditional Independence

We introduced earlier the meaning of conditional probability, and we showed that the set of conditional probabilities $\{\Pr(A|B)\}$ specifies a valid probability law. It is therefore possible to discuss independence with respect to conditional probability. We say that events A_1 and A_2 are *conditionally independent*, given event B, if

$$\Pr(A_1 \cap A_2|B) = \Pr(A_1|B)\Pr(A_2|B).$$

Note that if A_1 and A_2 are conditionally independent, we can use equation (4.2) to write

$$\begin{aligned}\Pr(A_1 \cap A_2|B) &= \frac{\Pr(A_1 \cap A_2 \cap B)}{\Pr(B)}\\ &= \frac{\Pr(B)\Pr(A_1|B)\Pr(A_2|A_1 \cap B)}{\Pr(B)}\\ &= \Pr(A_1|B)\Pr(A_2|A_1 \cap B).\end{aligned}$$

Under the assumption that $\Pr(A_1|B) > 0$, we can combine the previous two expressions and get

$$\Pr(A_2|A_1 \cap B) = \Pr(A_2|B).$$

This latter result asserts that, given event B has taken place, the additional information that A_1 has also occurred does not affect the likelihood of A_2. It is simple to show that conditional independence is a symmetric relation as well.

Example 31. *Suppose that a fair coin is tossed until heads is observed. The number of trials is recorded as the outcome of this experiment. We denote by B the event that the coin is tossed more than one time. Moreover, we let A_1 be the event that the number of trials is an even number; and A_2, the event that the number of trials is less than six. The conditional probabilities of A_1 and A_2, given that the coin is tossed more than once, are*

$$\begin{aligned}\Pr(A_1|B) &= \frac{\Pr(A_1 \cap B)}{\Pr(B)} = \frac{1/3}{1/2} = \frac{2}{3} \\ \Pr(A_2|B) &= \frac{\Pr(A_2 \cap B)}{\Pr(B)} = \frac{15/32}{1/2} = \frac{15}{16}.\end{aligned}$$

The joint probability of events A_1 and A_2 given B is equal to

$$\begin{aligned}\Pr(A_1 \cap A_2|B) &= \frac{\Pr(A_1 \cap A_2 \cap B)}{\Pr(B)} \\ &= \frac{5/16}{1/2} = \frac{5}{8} = \frac{2}{3} \cdot \frac{15}{16} \\ &= \Pr(A_1|B)\Pr(A_2|B).\end{aligned}$$

We conclude that A_1 and A_2 are conditionally independent given B. In particular, we have

$$\begin{aligned}\Pr(A_2|A_1 \cap B) &= \Pr(A_2|B) \\ \Pr(A_1|A_2 \cap B) &= \Pr(A_1|B).\end{aligned}$$

We emphasize that events A_1 and A_2 are not independent with respect to the unconditional probability law.

Two events that are independent with respect to an unconditional probability law may not be conditionally independent.

Example 32. *Two dice are rolled at the same time, a red die and a blue die. We can easily compute the probability of simultaneously getting a two on the red die and a six on the blue die,*

$$\Pr(\{r = 2\} \cap \{b = 6\}) = \frac{1}{36} = \Pr(r = 2) \Pr(b = 6).$$

Clearly, these two events are independent.

Consider the probability of rolling a two on the red die and a six on the blue die given that the sum of the two dice is an odd number. The individual conditional probabilities are given by

$$\Pr(\{r = 2\} | \{r + b \textit{ is odd}\}) = \Pr(\{b = 6\} | \{r + b \textit{ is odd}\}) = \frac{1}{6},$$

whereas the joint conditional probability is

$$\Pr(\{r = 2\} \cap \{b = 6\} | \{r + b \textit{ is odd}\}) = 0.$$

These two events are not conditionally independent.

It is possible to extend the notion of conditional independence to several events. The events $A_1, A_2, \ldots, A_n$ are conditionally independent given B if

$$\Pr\left(\bigcap_{i \in \mathbb{I}} A_i \middle| B \right) = \prod_{i \in \mathbb{I}} \Pr(A_i | B)$$

for every subset $\mathbb{I}$ of $\{1, 2, \ldots, n\}$. This definition is analogous to (4.6), albeit using the appropriate conditional probability law.

4.5 Equivalent Notations

In the study of probability, we are frequently interested in the probability of multiple events occurring simultaneously. So far, we have expressed the joint probability of events A and B using the notation $\Pr(A \cap B)$. For mathematical convenience, we also represent the probability that two events occur at the same time by

$$\Pr(A, B) = \Pr(A \cap B).$$

This alternate notation easily extends to the joint probability of several events. We denote the joint probability of events $A_1, A_2, \ldots, A_n$ by

$$\Pr(A_1, A_2, \ldots, A_n) = \Pr\left(\bigcap_{k=1}^{n} A_k \right).$$

Conditional probabilities can be written using a similar format. The probability of A given events $B_1, B_2, \ldots, B_n$ becomes

$$\Pr(A | B_1, B_2, \ldots, B_n) = \Pr\left(A \middle| \bigcap_{k=1}^{n} B_k \right).$$

From this point forward, we use these equivalent notations interchangeably.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Chapter 3.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Sections 1.3–1.5.

3. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Sections 2.4–2.6.

4. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 1.5–1.6.

Chapter 5

Discrete Random Variables

Suppose that an experiment and a sample space are given. A *random variable* is a real-valued function of the outcome of the experiment. In other words, the random variable assigns a specific number to every possible outcome of the experiment. The numerical value of a particular outcome is simply called the *value* of the random variable. Because of the structure of real numbers, it is possible to define pertinent statistical properties on random variables that otherwise do not apply to probability spaces in general.

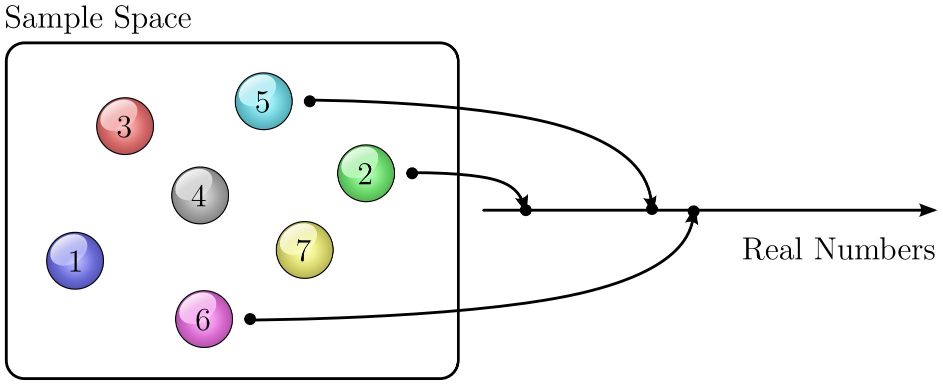

Figure 5.1: The sample space in this example has seven possible outcomes. A random variable maps each of these outcomes to a real number.

Example 33. *There are six possible outcomes to the rolling of a fair die, namely each of the six faces. These faces map naturally to the integers one* *through six. The value of the random variable, in this case, is simply the number of dots that appear on the top face of the die.*

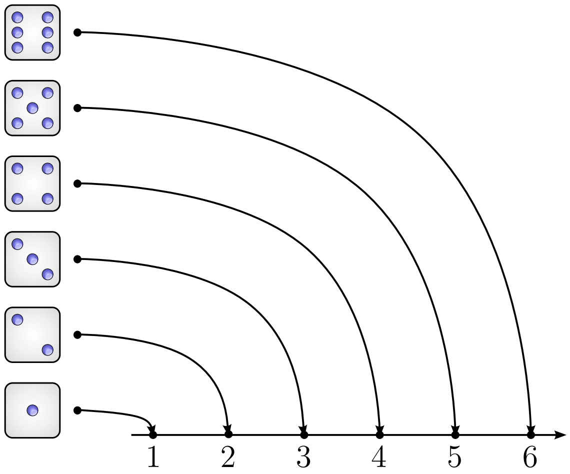

Figure 5.2: This random variable takes its input from the rolling of a die and assigns to each outcome a real number that corresponds to the number of dots that appear on the top face of the die.

A simple class of random variables is the collection of *discrete random variables.* A variable is called discrete if its range is finite or countably infinite; that is, it can only take a finite or countable number of values.

Example 34. *Consider the experiment where a coin is tossed repetitively until heads is observed. The corresponding function, which maps the number of tosses to an integer, is a discrete random variable that takes a countable number of values. The range of this random variable is given by the positive integers* $\{1, 2, \ldots\}$.

5.1 Probability Mass Functions

A discrete random variable X is characterized by the probability of each of the elements in its range. We identify the probabilities of individual elements in the range of X using the *probability mass function (PMF)* of X, which we denote by $p_X(\cdot)$. If x is a possible value of X then the *probability mass* of x, written $p_X(x)$, is defined by

$$p_X(x) = \Pr(\{X = x\}) = \Pr(X = x). \tag{5.1}$$

Equivalently, we can think of $p_X(x)$ as the probability of the set of all outcomes in Ω for which X is equal to x,

$$p_X(x) = \Pr(X^{-1}(x)) = \Pr(\{\omega \in \Omega | X(\omega) = x\}).$$

Here, $X^{-1}(x)$ denotes the *preimage* of x defined by $\{\omega \in \Omega | X(\omega) = x\}$. This is not to be confused with the inverse of a bijection.

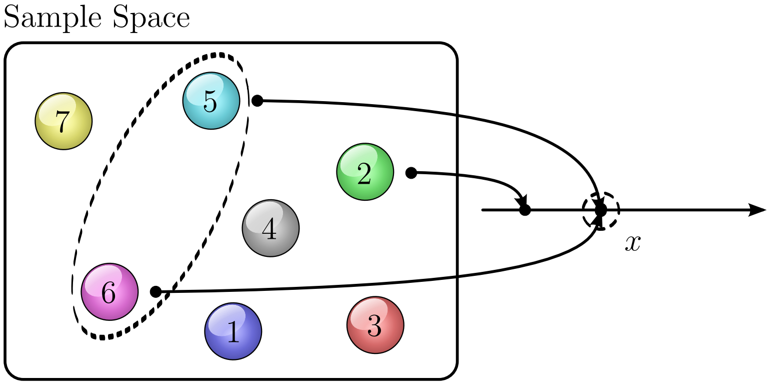

Figure 5.3: The probability mass of x is given by the probability of the set of all outcomes which X maps to x.

Let $X(\Omega)$ denote the collection of all the possible numerical values X can take; this set is known as the range of X. Using this notation, we can write

$$\sum_{x \in X(\Omega)} p_X(x) = 1. \tag{5.2}$$

We emphasize that the sets defined by $\{\omega \in \Omega | X(\omega) = x\}$ are disjoint and form a partition of the sample space Ω, as x ranges over all the possible values in $X(\Omega)$. Thus, (5.2) follows immediately from the countable additivity axiom and the normalization axiom of probability laws. In general, if X is a discrete random variable and S is a subset of $X(\Omega)$, we can write

$$\Pr(S) = \Pr\left(\{\omega \in \Omega | X(\omega) \in S\}\right) = \sum_{x \in S} p_X(x). \tag{5.3}$$

This equation offers an explicit formula to compute the probability of any subset of $X(\Omega)$, provided that X is discrete.

Example 35. *An urn contains three balls numbered one, two and three. Two balls are drawn from the urn without replacement. We wish to find the probability that the sum of the two selected numbers is odd.*

Let Ω be the set of ordered pairs corresponding to the possible outcomes of the experiment, $\Omega = \{(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)\}$. Note that these outcomes are equiprobable. We employ X to represent the sum of the two selected numbers. The PMF of random variable X is given by

$$p_X(3) = p_X(4) = p_X(5) = \frac{1}{3}.$$

If S denotes the event that the sum of the two numbers is odd, then the probability of the sum being odd can be computed as follows,

$$\Pr(S) = \Pr(\{3,5\}) = p_X(3) + p_X(5) = \frac{2}{3}.$$

5.2 Important Discrete Random Variables

A number of discrete random variables appears frequently in problems related to probability. These random variables arise in many different contexts, and they are worth getting acquainted with. In general, discrete random variables occur primarily in situations where counting is involved.

5.2.1 Bernoulli Random Variables

The first and simplest random variable is the *Bernoulli random variable*. Let X be a random variable that takes on only two possible numerical values, $X(\Omega) = \{0,1\}$. Then, X is a Bernoulli random variable and its PMF is given by

$$p_X(x) = \begin{cases} 1-p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \end{cases}$$

where $p \in [0,1]$.

Example 36. *Consider the flipping of a biased coin, for which heads is obtained with probability p and tails is obtained with probability $1-p$. A random variable that maps heads to one and tails to zero is a Bernoulli random variable with parameter p. In fact, every Bernoulli random variable is equivalent to the tossing of a coin.*

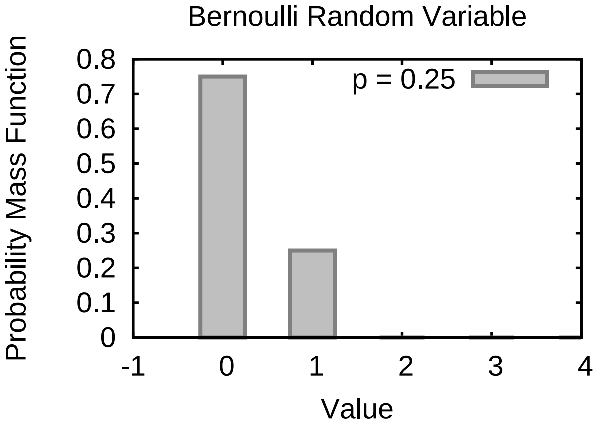

Figure 5.4: The PMF of a Bernoulli random variable appears above for parameter $p = 0.25$.

5.2.2 Binomial Random Variables

Multiple independent Bernoulli random variables can be combined to construct more sophisticated random variables. Suppose X is the sum of n independent and identically distributed Bernoulli random variables. Then X is called a *binomial random variable* with parameters n and p. The PMF of X is given by

$$p_X(k) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $k = 0, 1, \ldots n$. We can easily verify that X fulfills the normalization axiom,

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

Example 37. *The Brazos Soda Company creates an "Under the Caps" promotion whereby a customer can win an instant cash prize of $1 by looking under a bottle cap. The likelihood to win is one in four, and it is independent from bottle to bottle. A customer buys eight bottles of soda from this company. We wish to find the PMF of the number of winning bottles, which we denote by X. Also, we want to compute the probability of winning more than $4.*

The random variable X is binomial with parameters $n = 8$ and $p = 1/4$.

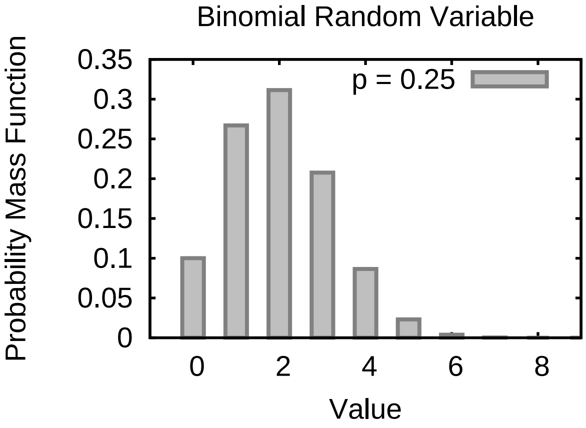

Figure 5.5: This figure shows a binomial random variable with parameters $n = 8$ and $p = 0.25$.

Its PMF is given by

$$p_X(k) = \binom{8}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{8-k} = \binom{8}{k} \frac{3^{8-k}}{4^8},$$

where $k = 0, 1, \ldots, 8$. The probability of winning more than \$4 is

$$\Pr(X > 4) = \sum_{k=5}^{8} \binom{8}{k} \frac{3^{8-k}}{4^8}.$$

5.2.3 Poisson Random Variables

The probability mass function of a *Poisson random variable* is given by

$$p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots$$

where λ is a positive number. Note that, using Taylor series expansion, we have

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1,$$

which shows that this PMF fulfills the normalization axiom of probability laws. The Poisson random variable is of fundamental importance when counting the number of occurrences of a phenomenon within a certain time period. It finds extensive use in networking, inventory management and queueing applications.

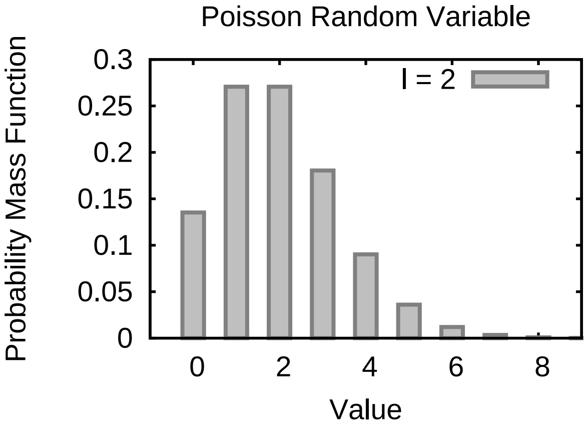

Figure 5.6: This figure shows the PMF of a Poisson random variable with parameter $\lambda = 2$. Note that the values of the PMF are only present for $k = 0, 1, \ldots, 8$.

Example 38. *Requests at an Internet server arrive at a rate of λ connections per second. The number of service requests per second is modeled as a random variable with a Poisson distribution. We wish to find the probability that no service requests arrive during a time interval of one second.*

Let N be a random variable that represents the number of requests that arrives within a span of one second. By assumption, N is a Poisson random variable with PMF

$$p_N(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The probability that no service requests arrive in one second is simply given by $p_N(0) = e^{-\lambda}$.

It is possible to obtain a Poisson random variable as the limit of a sequence of binomial random variables. Fix λ and let $p_n = \lambda/n$. For $k = 1, 2, \ldots n$, we define the PMF of the random variable X_n as

$$\begin{aligned} p_{X_n}(k) = \Pr(X_n = k) &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n} \right)^k \left(1 - \frac{\lambda}{n} \right)^{n-k} \\ &= \frac{n!}{n^k (n-k)!} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n} \right)^{n-k}. \end{aligned}$$

In the limit, as n approaches infinity, we get

$$\lim_{n\to\infty} p_{X_n}(k) = \lim_{n\to\infty} \frac{n!}{n^k(n-k)!} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Thus, the sequence of binomial random variables $\{X_n\}$ converges in distribution to the Poisson random variable X.

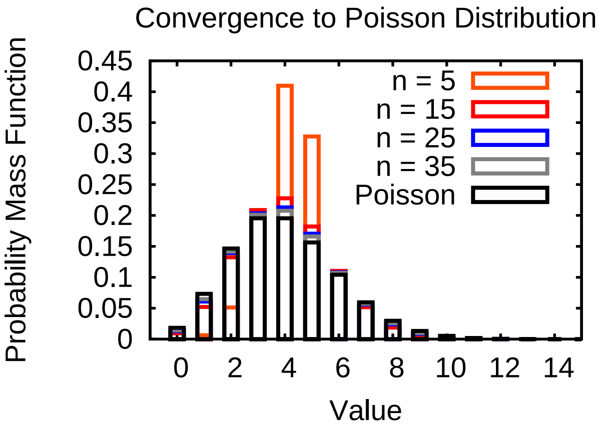

Figure 5.7: The levels of a binomial PMF with parameter $p = \lambda/n$ converge to the probabilities of a Poisson PMF with parameter λ as n increases to infinity.

This discussion suggests that the Poisson PMF can be employed to approximate the PMF of a binomial random variable in certain circumstances. Suppose that Y is a binomial random variable with parameters n and p. If n is large and p is small then the probability that Y equals k can be approximated by

$$p_Y(k) = \frac{n!}{n^k(n-k)!} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda},$$

where $\lambda = np$. The latter expression can be computed numerically in a straightforward manner.

Example 39. *The probability of a bit error on a communication channel is equal to* 10^{-2}*. We wish to approximate the probability that a block of* 1000 *bits has four or more errors.*

Assume that the probability of individual errors is independent from bit to bit. The transmission of each bit can be modeled as a Bernoulli trial, with a *zero indicating a correct transmission and a one representing a bit error. The total number of errors in* 1000 *transmissions then corresponds to a binomial random variable with parameters* $n = 1000$ *and* $p = 10^{-2}$. *The probability of making four or more errors can be approximated using a Poisson random variable with constant* $\lambda = np = 10$. *Thus, we can approximate the probability that a block of* 1000 *bits has four or more errors by*

$$\begin{aligned}\Pr(N \geq 4) &= 1 - \Pr(N < 4) \approx 1 - \sum_{k=0}^{3} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= 1 - e^{-10} \left(1 + 10 + 50 + \frac{500}{3} \right) \\ &\approx 0.9897.\end{aligned}$$

This is in contrast to the exact answer, which can be written as 0.9899 *when truncated to four decimal places.*

5.2.4 Geometric Random Variables

Consider a random experiment where a Bernoulli trial is repeated multiple times until a one is observed. At each time step, the probability of getting a one is equal to p and the probability of getting a zero is $1 - p$. The number of trials carried out before completion, which we denote by X, is recorded as the outcome of this experiment. The random variable X is a *geometric random variable*, and its PMF is given by

$$p_X(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \ldots$$

We stress that $(1 - p)^{k-1} p$ simply represents the probability of obtaining a sequence of $k - 1$ zero immediately followed by a one.

Example 40. *The Brazos Soda Company introduces another "Under the Caps" promotion. This time, a customer can win an additional bottle of soda by looking under the cap of her bottle. The probability to win is* $1/5$*, and it is independent from bottle to bottle. A customer purchases one bottle of soda from the Brazos Soda Company and thereby enters the contest. For every extra bottle of soda won by looking under the cap, the customer gets an additional chance*

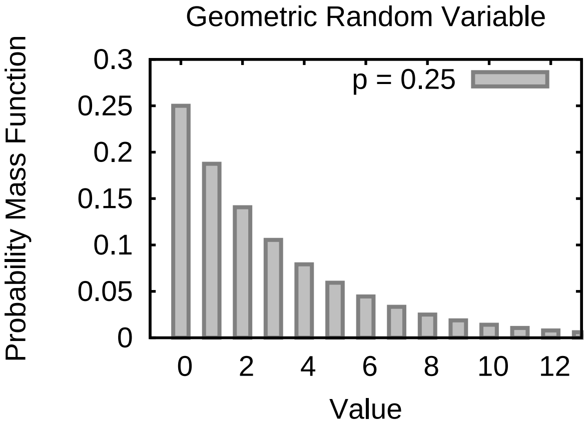

Figure 5.8: The PMF of a geometric random variable is a decreasing function of k. It is plotted above for $p = 0.25$. The values of the PMF are only present for $k = 1, 2, \ldots, 12$.

to play. We wish to find the PMF of the number of bottles obtained by this customer.

Let X denote the total number of bottles obtained by the customer. The random variable X is geometric and its PMF is

$$p_X(k) = \left(\frac{1}{5}\right)^{k-1} \frac{4}{5},$$

where $k = 1, 2, \ldots$

Memoryless Property: A remarkable aspect of the geometric random variable is that it satisfies the *memoryless property*,

$$\Pr(X = k + j | X > k) = \Pr(X = j).$$

This can be established using the definition of conditional probability. Let X be a geometric random variable with parameter p, and assume that k and j are positive integers. We can write the conditional probability of X as

$$\begin{aligned}\Pr(X = k + j | X > k) &= \frac{\Pr(\{X = k + j\} \cap \{X > k\})}{\Pr(X > k)} \\ &= \frac{\Pr(X = k + j)}{\Pr(X > k)} = \frac{(1-p)^{k+j-1} p}{(1-p)^k} \\ &= (1-p)^{j-1} p = \Pr(X = j).\end{aligned}$$

In words, the probability that the number of trials carried out before completion is $k + j$, given k unsuccessful trials, is equal to the unconditioned probability that the total number of trials is j. It can be shown that the geometric random variable is the only discrete random variable that possesses the memoryless property.

5.2.5 Discrete Uniform Random Variables

A finite random variable where all the possible outcomes are equally likely is called a *discrete uniform random variable*. Let X be a uniform random variable taking value over $X(\Omega) = \{1, 2, \ldots, n\}$. Its PMF is therefore given by

$$p_X(k) = \begin{cases} 1/n, & \text{if } k = 1, 2, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

We encountered specific cases of this random variable before. The tossing of a fair coin and the rolling of a die can both be used to construct uniform random variables.

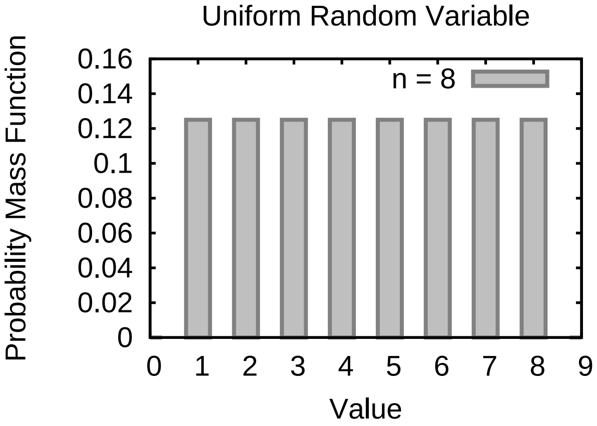

Figure 5.9: A uniform random variable corresponds to the situation where all the possible values of the random variable are equally likely. It is shown above for the case where $n = 8$.

5.3 Functions of Random Variables

Recall that a random variable is a function of the outcome of an experiment. Given a random variable X, it is possible to create a new random variable Y by applying a real-valued function $g(\cdot)$ to X. If X is a random variable then $Y = g(X)$ is also a random variable since it associates a numerical value to every outcome of the experiment. In particular, if $\omega \in \Omega$ is the outcome of the experiment, then X takes on value $x = X(\omega)$ and Y takes on value $Y(\omega) = g(x)$.

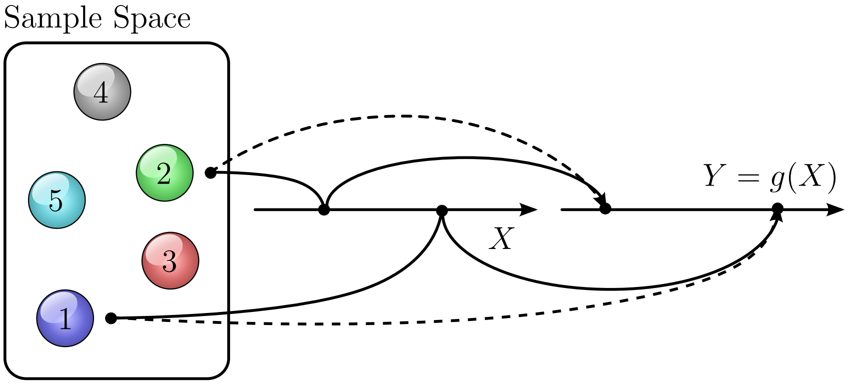

Figure 5.10: A real-valued function of a random variable is a random variable itself. In this figure, Y is obtained by passing random variable X through a function $g(\cdot)$.

Furthermore, if X is a discrete random variable, then so is Y. The set of possible values Y can take is denoted by $g(X(\Omega))$, and the number of elements in $g(X(\Omega))$ is no greater than the number of elements in $X(\Omega)$. The PMF of Y, which we represent by $p_Y(\cdot)$, is obtained as follows. If $y \in g(X(\Omega))$ then

$$p_Y(y) = \sum_{\{x \in X(\Omega) | g(x) = y\}} p_X(x); \tag{5.4}$$

otherwise, $p_Y(y) = 0$. In particular, $p_Y(y)$ is non-zero only if there exists an $x \in X(\Omega)$ such that $g(x) = y$ and $p_X(x) > 0$.

Example 41. *Let X be a random variable and let $Y = g(X) = aX + b$, where a and b are constant. That is, Y is an affine function of X. Suppose that*

$a \neq 0$, then the probability of Y being equal to value y is given by

$$p_Y(y) = p_X\left(\frac{y-b}{a}\right).$$

Linear and affine functions of random variables are commonplace in applied probability and engineering.

Example 42. *A taxi driver works in New York City. The distance of a ride with a typical client is a discrete uniform random variable taking value over $X(\Omega) = \{1, 2, \ldots, 10\}$. The metered rate of fare, according to city rules, is \$2.50 upon entry and \$0.40 for each additional unit (one-fifth of a mile). We wish to find the PMF of Y, the value of a typical fare.*

Traveling one mile is five units and costs an extra \$2.00. The smallest possible fare is therefore \$4.50, with probability $1/10$. Similarly, a ride of X miles will generate a revenue of $Y = 2.5 + 2X$ dollars for the cab driver. The PMF of Y is thus given by

$$p_Y(2.5 + 2k) = \frac{1}{10}$$

for $k = 1, 2, \ldots, 10$; and it is necessarily zero for any other argument.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Sections 4.1–4.2, 4.6–4.8.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Sections 2.1–2.3.

3. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Section 2.7.

4. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 2.1–1.2.

Chapter 6

Meeting Expectations

When a large collection of data is gathered, one is typically interested not necessarily in every individual data point, but rather in certain descriptive quantities such as the average or the median. The same is true for random variables. The PMF of discrete random variable X provides a complete characterization of the distribution of X by specifying the probability of every possible value of X. Still, it may be desirable to summarize the information contained in the PMF of a random variable. One way to accomplish this task and thereby obtain meaningful descriptive quantities is through the expectation operator.

6.1 Expected Values

The *expected value* $\mathrm{E}[X]$ of discrete random variable X is defined by

$$\mathrm{E}[X] = \sum_{x \in X(\Omega)} x p_X(x), \tag{6.1}$$

whenever this sum converges absolutely. If this sum is not absolutely convergent, then X is said not to possess an expected value. As mentioned above, the expected value $\mathrm{E}[X]$ provides insightful information about the underlying random variable X without giving a comprehensive and overly detailed description. The expected value of a random variable, as defined in (6.1), is also called the *mean* of X. It is important to realize that $\mathrm{E}[X]$ is not a function of random variable X; rather, it is a function of the PMF of X.

Example 43. *A fair die is rolled once, with the number of dots appearing on the top face taken as the value of the corresponding random variable. The expected value of the roll can be computed as*

$$\sum_{k=1}^{6} \frac{k}{6} = \frac{42}{12} = 3.5.$$

In other words, the mean of this random variable is 3.5*.*

Example 44. *Assume that a fair coin is flipped repetitively until heads is observed. The value of random variable X is taken to be the total number of tosses performed during this experiment. The possible values for X are therefore given by $X(\Omega) = \{1, 2, \ldots\}$. Recall that, in this case, the PMF of X is equal to $p_X(k) = 2^{-k}$, where k is a positive integer. The expected value of this geometric random variable can be computed as*

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} \frac{k}{2^k} = 2.$$

The expected number of tosses until the coin produces heads is equal to two.

In general, determining the expectation of a random variable requires as input its PMF, a detailed characterization of the random variable, and returns a much simpler scalar attribute, its mean. Hence, computing the expected value of the random variable yields a concise summary of its overall behavior.

6.2 Functions and Expectations

The mean forms one instance where the distribution of a random variable is condensed into a scalar quantity. There are several additional examples. The notion of an *expectation* can be combined with traditional functions to create alternate descriptions and other meaningful quantities. Suppose that X is a discrete random variable. Let $g(\cdot)$ be a real-valued function on the range of X, and consider the expectation of $g(X)$. This expected value, $\mathrm{E}[g(X)]$, is a scalar quantity that can also provide partial information about the distribution of X.

One way to determine the expected value of $g(X)$ is to first note that $Y = g(X)$ is itself a random variable. Thus, we can find the derived distribution of Y and then apply the definition of expected value provided in (6.1). Yet, there is a more direct way to compute this quantity; the *expectation* of $g(X)$ can be expressed as

$$\mathrm{E}\left[g(X)\right] = \sum_{x \in X(\Omega)} g(x) p_X(x). \tag{6.2}$$

It is worth re-emphasizing that there exist random variables and functions for which the above sum does not converge. In such cases, we simply say that the expected value of $g(X)$ does not exist. Also, notice that the mean $\mathrm{E}[X]$ is a special case of (6.2), where $g(X) = X$. Hence the definition of $\mathrm{E}[g(X)]$ given above subsumes our original description of $\mathrm{E}[X]$, which appeared in (6.1). We explore pertinent examples below.

Example 45. *The simplest possible scenario for* (6.2) *is the case where the function $g(\cdot)$ is a constant. The expectation of $g(X) = c$ becomes*

$$\mathrm{E}[c] = \sum_{x \in X(\Omega)} c p_X(x) = c \sum_{x \in X(\Omega)} p_X(x) = c.$$

The last inequality follows from the normalization axiom of probability laws. The expectation of a constant is always the constant itself.

Example 46. *Let S be a subset of the real numbers, and define the* indicator function *of S by*

$$\mathbf{1}_S(x) = \begin{cases} 1, & x \in S \\ 0, & x \notin S. \end{cases}$$

The expectation of $\mathbf{1}_S(X)$ is equal to

$$\begin{aligned} \mathrm{E}\left[\mathbf{1}_S(X)\right] &= \sum_{x \in X(\Omega)} \mathbf{1}_S(x) p_X(x) \\ &= \sum_{x \in S \cap X(\Omega)} p_X(x) = \Pr(X \in S). \end{aligned}$$

That is, the expectation of the indicator function of S is simply the probability that X takes on a value in S. This alternate way of computing the probability of an event can sometimes be employed to solve otherwise difficult probability problems.

Let random variable Y be defined by applying real-valued function $g(\cdot)$ to X, with $Y = g(X)$. The mean of Y is equal to the expectation of $g(X)$, and we know from our ongoing discussion that this value can be obtained by applying two different formulas. To ensure consistency, we verify that these two approaches lead to a same answer.

First, we can apply (6.1) directly to Y, and obtain

$$\mathrm{E}[Y] = \sum_{y \in g(X(\Omega))} y p_Y(y),$$

where $p_Y(\cdot)$ is the PMF of Y provided by (5.3). Alternatively, using (6.2), we have

$$\mathrm{E}[Y] = \mathrm{E}[g(X)] = \sum_{x \in X(\Omega)} g(x) p_X(x).$$

We prove that these two expressions describe a same answer as follows. Recall that the PMF of Y evaluated at y is obtained by summing the values of $p_X(\cdot)$ over all $x \in X(\Omega)$ such that $g(x) = y$. Mathematically, this can be expressed as $p_Y(y) = \sum_{\{x \in X(\Omega) | g(x) = y\}} p_X(x)$. Using this equality, we can write

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_{y \in g(X(\Omega))} y p_Y(y) = \sum_{y \in g(X(\Omega))} y \sum_{\{x \in X(\Omega) | g(x) = y\}} p_X(x) \\
&= \sum_{y \in g(X(\Omega))} \sum_{\{x \in X(\Omega) | g(x) = y\}} y p_X(x) \\
&= \sum_{y \in g(X(\Omega))} \sum_{\{x \in X(\Omega) | g(x) = y\}} g(x) p_X(x) \\
&= \sum_{x \in X(\Omega)} g(x) p_X(x) = \mathrm{E}[g(X)].
\end{aligned}$$

Note that first summing over all possible values of Y and then over the preimage of every $y \in Y(\Omega)$ is equivalent to summing over all $x \in X(\Omega)$. Hence, we have shown that computing the expectation of a function using the two methods outlined above leads to a same solution.

Example 47. *Brazos Extreme Events Radio creates the "Extreme Trio" contest. To participate, a person must fill out an application card. Three cards are drawn from the lot and each winner is awarded \$1,000. While a same participant can send multiple cards, he or she can only win one grand prize.*

At the time of the drawing, the radio station has accumulated a total of 100 cards. David, an over-enthusiastic listener, is accountable for half of these cards. We wish to compute the amount of money David expects to win under this promotion.

Let X be the number of cards drawn by the radio station written by David. The PMF of X is given by

$$p_X(k) = \frac{\binom{50}{k}\binom{50}{3-k}}{\binom{100}{3}} \quad k \in \{0, 1, 2, 3\}.$$

The money earned by David can be expressed as $g(k) = 1000 \min\{k, 1\}$. It follows that the expected amount of money he receives is equal to

$$\sum_{k=0}^{3} (1000 \min\{k, 1\})\, p_X(k) = 1000 \cdot \frac{29}{33}.$$

Alternatively, we can define $Y = 1000 \min\{X, 1\}$. Clearly, Y can only take on one of two possible values, 0 or 1000. Evaluating the PMF of Y, we get $p_Y(0) = p_X(0) = 4/33$ and $p_Y(1000) = 1 - p_Y(0) = 29/33$. The expected value of Y is equal to

$$0 \cdot p_Y(0) + 1000 \cdot p_Y(1000) = 1000 \cdot \frac{29}{33}.$$

As anticipated, both methods lead to the same answer. The expected amount of money won by David is roughly $878.79.

6.2.1 The Mean

As seen at the beginning of this chapter, the simplest non-trivial expectation is the *mean*. We provide two additional examples for the mean, and we explore a physical interpretation of its definition below.

Example 48. *Let X be a geometric random variable with parameter p and PMF*

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots$$

The mean of this random variable is

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = p \sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{1}{p}.$$

Example 49. *Let X be a binomial random variable with parameters n and p. The PMF of X is given by*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

The mean of this binomial random variable can therefore be computed as

$$\begin{aligned} \mathrm{E}[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{\ell=0}^{n-1} \frac{n!}{\ell!(n-1-\ell)!} p^{\ell+1} (1-p)^{n-\ell-1} \\ &= np \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p^{\ell} (1-p)^{n-1-\ell} = np. \end{aligned}$$

Notice how we rearranged the sum into a familiar form to compute its value.

It can be insightful to relate the mean of a random variable to classical mechanics. Let X be a random variable and suppose that, for every $x \in X(\Omega)$, we place an infinitesimal particle of mass $p_X(x)$ at position x along a real line. The mean of random variable X as defined in (6.1) coincides with the center of mass of the system of particles.

Example 50. *Let X be a Bernoulli random variable such that*

$$p_X(x) = \begin{cases} 0.25, & \text{if } x = 0 \\ 0.75, & \text{if } x = 1. \end{cases}$$

The mean of X is given by

$$\mathrm{E}[X] = 0 \cdot 0.25 + 1 \cdot 0.75 = 0.75.$$

Consider a two-particle system with masses $m_1 = 0.25$ and $m_2 = 0.75$, respectively. In the coordinate system illustrated below, the particles are located at positions $x_1 = 0$ and $x_2 = 1$. From classical mechanics, we know that their center of mass can be expressed as

$$\frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} = 0.75.$$

As anticipated, the center of mass corresponds to the mean of X.

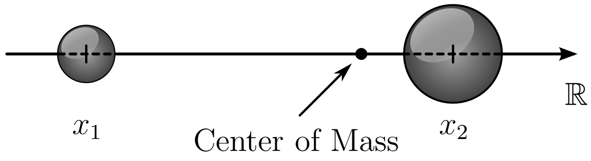

Figure 6.1: The center of mass on the figure is indicated by the tip of the arrow. In general, the mean of a discrete random variable corresponds to the center of mass of the associated particle system

6.2.2 The Variance

A second widespread descriptive quantity associated with random variable X is its *variance*, which we denote by $\mathrm{Var}[X]$. It is defined by

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2 \right]. \tag{6.3}$$

Evidently, the variance is always nonnegative. It provides a measure of the dispersion of X around its mean. For discrete random variables, it can be computed explicitly as

$$\mathrm{Var}[X] = \sum_{x \in X(\Omega)} (x - \mathrm{E}[X])^2 \, p_X(x).$$

The square root of the variance is referred to as the *standard deviation* of X, and it is often denoted by σ.

Example 51. *Suppose X is a Bernoulli random variable with parameter p. We can compute the mean of X as*

$$\mathrm{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p.$$

Its variance is given by

$$\mathrm{Var}[X] = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p).$$

Example 52. *Let X be a Poisson random variable with parameter λ. The mean of X is given by*

$$\begin{aligned} \mathrm{E}[X] &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\ &= \lambda \sum_{\ell=0}^{\infty} \frac{\lambda^{\ell}}{\ell!} e^{-\lambda} = \lambda. \end{aligned}$$

The variance of X can be calculated as

$$\begin{aligned}\mathrm{Var}[X] &= \sum_{k=0}^{\infty} (k-\lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \left(\lambda^2 + k(1-2\lambda) + k(k-1)\right) \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda - \lambda^2 + \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda} = \lambda.\end{aligned}$$

Both the mean and the variance of a Poisson random variable are equal to its parameter λ.

6.2.3 Affine Functions

Proposition 3. *Suppose X is a random variable with finite mean. Let Y be the affine function of X defined by $Y = aX + b$, where a and b are fixed real numbers. The mean of random variable Y is equal to* $\mathrm{E}[Y] = a\mathrm{E}[X] + b$.

Proof. This can be computed using (6.2);

$$\begin{aligned}\mathrm{E}[Y] &= \sum_{x \in X(\Omega)} (ax+b) p_X(x) \\ &= a \sum_{x \in X(\Omega)} x p_X(x) + b \sum_{x \in X(\Omega)} p_X(x) \\ &= a\mathrm{E}[X] + b.\end{aligned}$$

We can summarize this property with $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$. □

It is not much harder to show that the expectation is a linear functional. Suppose that $g(\cdot)$ and $h(\cdot)$ are two real-valued functions such that $\mathrm{E}[g(X)]$ and $\mathrm{E}[h(X)]$ both exist. We can write the expectation of $ag(X) + h(X)$ as

$$\begin{aligned}\mathrm{E}[ag(X) + h(X)] &= \sum_{x \in X(\Omega)} (ag(x) + h(x)) p_X(x) \\ &= a \sum_{x \in X(\Omega)} g(x) p_X(x) + \sum_{x \in X(\Omega)} h(x) p_X(x) \\ &= a\mathrm{E}[g(X)] + \mathrm{E}[h(X)].\end{aligned}$$

This demonstrates that the expectation is both homogeneous and additive.

Proposition 4. *Assume that X is a random variable with finite mean and variance, and let Y be the affine function of X given by $Y = aX + b$, where a and b are constants. The variance of Y is equal to* $\mathrm{Var}[Y] = a^2 \mathrm{Var}[X]$.

Proof. Consider (6.3) applied to $Y = aX + b$,

$$\begin{aligned}\mathrm{Var}[Y] &= \sum_{x \in X(\Omega)} \left(ax + b - \mathrm{E}[aX + b] \right)^2 p_X(x) \\ &= \sum_{x \in X(\Omega)} \left(ax + b - a\mathrm{E}[X] - b \right)^2 p_X(x) \\ &= a^2 \sum_{x \in X(\Omega)} \left(x - \mathrm{E}[X] \right)^2 p_X(x) = a^2 \mathrm{Var}[X].\end{aligned}$$

The variance of an affine function only depends on the distribution of its argument and parameter a. A translation of the argument by b does not affect the variance of $Y = aX + b$; in other words, it is shift invariant. □

6.3 Moments

The *moments* of a random variable X are likewise important quantities used in providing partial information about the PMF of X. The nth moment of random variable X is defined by

$$\mathrm{E}[X^n] = \sum_{x \in X(\Omega)} x^n p_X(x). \tag{6.4}$$

Incidentally, the mean of random variable X is its first moment.

Proposition 5. *The variance of random variable X can be expressed in terms of its first two moments,* $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2$.

Proof. Suppose that the variance of X exists and is finite. Starting from (6.3), we can expand the variance of X as follows,

$$\begin{aligned}\mathrm{Var}[X] &= \sum_{x \in X(\Omega)} \left(x - \mathrm{E}[X] \right)^2 p_X(x) \\ &= \sum_{x \in X(\Omega)} \left(x^2 - 2x\mathrm{E}[X] + (\mathrm{E}[X])^2 \right) p_X(x) \\ &= \sum_{x \in X(\Omega)} x^2 p_X(x) - 2\mathrm{E}[X] \sum_{x \in X(\Omega)} x p_X(x) + (\mathrm{E}[X])^2 \sum_{x \in X(\Omega)} p_X(x) \\ &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 .\end{aligned}$$

This alternate formula for the variance is sometimes convenient for computational purposes. □

We include below an example where the above formula for the variance is applied. This allows the straightforward application of standard sums from calculus.

Example 53. *Let X be a uniform random variable with PMF*

$$p_X(k) = \begin{cases} 1/n, & \text{if } k = 1, 2, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

The mean of this uniform random variable is equal to

$$\mathrm{E}[X] = \frac{n+1}{2}.$$

The variance of X can then be obtained as

$$\begin{aligned} \mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 &= \sum_{k=1}^{n} \frac{k^2}{n} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n^2-1}{12}. \end{aligned}$$

Closely related to the moments of a random variable are its *central moments*. The kth central moment of X is defined by $\mathrm{E}\left[(X - E[X])^k\right]$. The variance is an example of a central moment, as we can see from definition (6.3). The central moments are used to define the *skewness* of random variable, which is a measure of asymmetry; and its *kurtosis*, which assesses whether the variance is due to infrequent extreme deviations or more frequent, modest-size deviations. Although these quantities will not play a central role in our exposition of probability, they each reveal a different characteristic of a random variable and they are encountered frequently in statistics.

6.4 Ordinary Generating Functions

In the special yet important case where $X(\Omega)$ is a subset of the non-negative integers, it is occasionally useful to employ the *ordinary generating function.*

This function bears a close resemblance to the *z-transform* and is defined by

$$G_X(z) = \mathrm{E}\left[z^X\right] = \sum_{k=0}^{\infty} z^k p_X(k). \tag{6.5}$$

It is also called the *probability-generating function* because of the following property. The probability $\Pr(X = k) = p_X(k)$ can be recovered from the corresponding generating function $G_X(z)$ through Taylor series expansion. Within the radius of convergence of $G_X(z)$, we have

$$G_X(z) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{d^k G_X}{dz^k}(0) \right) z^k.$$

Comparing this equation to (6.5), we conclude that

$$p_X(k) = \frac{1}{k!} \frac{d^k G_X}{dz^k}(0).$$

We note that, for $|z| \leq 1$, we have

$$\left| \sum_{k=0}^{\infty} z^k p_X(k) \right| \leq \sum_{k=0}^{\infty} |z|^k p_X(k) \leq \sum_{k=0}^{\infty} p_X(k) = 1$$

and hence the radius of convergence of any probability-generating function must include one.

The ordinary generating function plays an important role in dealing with sums of discrete random variables. As a preview of what lies ahead, we compute ordinary generating functions for Bernoulli and Binomial random variables below.

Example 54. *Let X be a Bernoulli random variable with parameter p. The ordinary generating function of X is given by*

$$G_X(z) = p_X(0) + p_X(1)z = 1 - p + pz.$$

Example 55. *Let S be a binomial random variable with parameters n and p. The ordinary generating function of S can be computed as*

$$\begin{aligned} G_S(z) &= \sum_{k=0}^{n} z^k p_S(k) = \sum_{k=0}^{n} z^k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pz)^k (1-p)^{n-k} = (1 - p + pz)^n. \end{aligned}$$

We know, from Section 5.2.2, that one way to create a binomial random variable is to sum n independent and identically distributed Bernoulli random variables, each with parameter p. Looking at the ordinary generating functions above, we notice that the $G_S(z)$ is the product of n copies of $G_X(z)$. Intuitively, it appears that the sum of independent discrete random variables leads to the product of their ordinary generating functions, a relation that we will revisit shortly.

The mean and second moment of X can be computed based on its ordinary generating function, In particular, we have

$$\mathrm{E}[X] = \lim_{z \uparrow 1} \frac{dG_X}{dz}(z).$$

Similarly, the second moment of X can be derived as

$$\mathrm{E}\left[X^2\right] = \lim_{z \uparrow 1} \left(\frac{d^2 G_X}{dz^2}(z) + \frac{dG_X}{dz}(z) \right).$$

This can be quite useful, as illustrated in the following example.

Example 56 (Poisson Random Variable)**.** *Suppose that X has a Poisson distribution with parameter $\lambda > 0$. The function $G_X(s)$ can be computed using the distribution of X,*

$$G_X(z) = \sum_{k=0}^{\infty} z^k \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda z)^k}{k!} = e^{-\lambda} e^{\lambda z} = e^{\lambda(z-1)}.$$

The first two moments of X are given by

$$\begin{aligned} \mathrm{E}[X] &= \lim_{z \uparrow 1} \frac{dG_X}{dz}(z) = \lim_{z \uparrow 1} \lambda e^{\lambda(z-1)} = \lambda \\ \mathrm{E}\left[X^2\right] &= \lim_{z \uparrow 1} \left(\frac{d^2 G_X}{dz^2}(z) + \frac{dG_X}{dz}(z) \right) = \lim_{z \uparrow 1} \left(\lambda^2 + \lambda\right) e^{\lambda(z-1)} = \lambda^2 + \lambda. \end{aligned}$$

This provides a very efficient way to compute the mean and variance of X, which are both equal to λ. It may be helpful to compare this derivation with Example 52.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Section 4.3.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 2.4.

3. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 2.4, 3.1.

4. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Sections 4.8.

Chapter 7

Multiple Discrete Random Variables

Thus far, our treatment of probability has been focused on single random variables. It is often convenient or required to model stochastic phenomena using multiple random variables. In this section, we extend some of the concepts developed for single random variables to multiple random variables. We center our exposition of random vectors around the simplest case, pairs of random variables.

7.1 Joint Probability Mass Functions

Consider two discrete random variables X and Y associated with a single experiment. The random pair (X, Y) is characterized by the *joint probability mass function* of X and Y, which we denote by $p_{X,Y}(\cdot, \cdot)$. If x is a possible value of X and y is a possible value of Y, then the probability mass function of (x, y) is denoted by

$$\begin{aligned} p_{X,Y}(x, y) &= \Pr(\{X = x\} \cap \{Y = y\}) \\ &= \Pr(X = x, Y = y). \end{aligned}$$

Note the similarity between the definition of the joint PMF and (5.1).

Suppose that S is a subset of $X(\Omega) \times Y(\Omega)$. We can express the probability

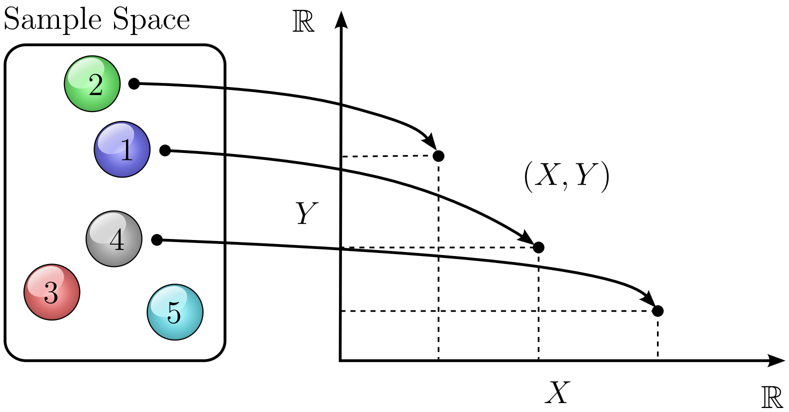

Figure 7.1: The random pair (X, Y) maps every outcome contained in the sample space to a vector in $\mathbb{R}^2$.

of S as

$$\begin{aligned} \Pr(S) &= \Pr(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in S\}) \\ &= \sum_{(x,y) \in S} p_{X,Y}(x, y). \end{aligned}$$

In particular, we have

$$\sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} p_{X,Y}(x, y) = 1.$$

To further distinguish between the joint PMF of (X, Y) and the individual PMFs $p_X(\cdot)$ and $p_Y(\cdot)$, we occasionally refer to the latter as *marginal probability mass functions*. We can compute the marginal PMFs of X and Y from the joint PMF $p_{X,Y}(\cdot, \cdot)$ using the formulas

$$\begin{aligned} p_X(x) &= \sum_{y \in Y(\Omega)} p_{X,Y}(x, y), \\ p_Y(y) &= \sum_{x \in X(\Omega)} p_{X,Y}(x, y). \end{aligned}$$

On the other hand, knowledge of the marginal distributions $p_X(\cdot)$ and $p_Y(\cdot)$ is not enough to obtain a complete description of the joint PMF $p_{X,Y}(\cdot, \cdot)$. This fact is illustrated in Examples 57 & 58.

Example 57. *An urn contains three balls numbered one, two and three. A random experiment consists of drawing two balls from the urn, without replacement. The number appearing on the first ball is a random variable, which we denote by X. Similarly, we refer to the number inscribed on the second ball as Y. The joint PMF of X and Y is specified in table form below,*

$p_{X,Y}(x,y)$	1	2	3
1	0	1/6	1/6
2	1/6	0	1/6
3	1/6	1/6	0

.

We can compute the marginal PMF of X as

$$p_X(x) = \sum_{y \in Y(\Omega)} p_{X,Y}(x,y) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3},$$

where $x \in \{1,2,3\}$. Likewise, the marginal PMF of Y is given by

$$p_Y(y) = \begin{cases} 1/3, & \textit{if } y \in \{1,2,3\} \\ 0, & \textit{otherwise.} \end{cases}$$

Example 58. *Again, suppose that an urn contains three balls numbered one, two and three. This time the random experiment consists of drawing two balls from the urn with replacement. We use X and Y to denote the numbers appearing on the first and second balls, respectively. The joint PMF of X and Y becomes*

$p_{X,Y}(x,y)$	1	2	3
1	1/9	1/9	1/9
2	1/9	1/9	1/9
3	1/9	1/9	1/9

.

The marginal distributions of X and Y are the same as in Example 57; however, the joint PMFs differ.

7.2 Functions and Expectations

Let X and Y be two random variables with joint PMF $p_{X,Y}(\cdot,\cdot)$. Consider a third random variable defined by $V = g(X,Y)$, where $g(\cdot,\cdot)$ is a real-valued function. We can obtain the PMF of V by computing

$$p_V(v) = \sum_{\{(x,y) | g(x,y)=v\}} p_{X,Y}(x,y). \tag{7.1}$$

This equation is the analog of (5.3) for pairs of random variables.

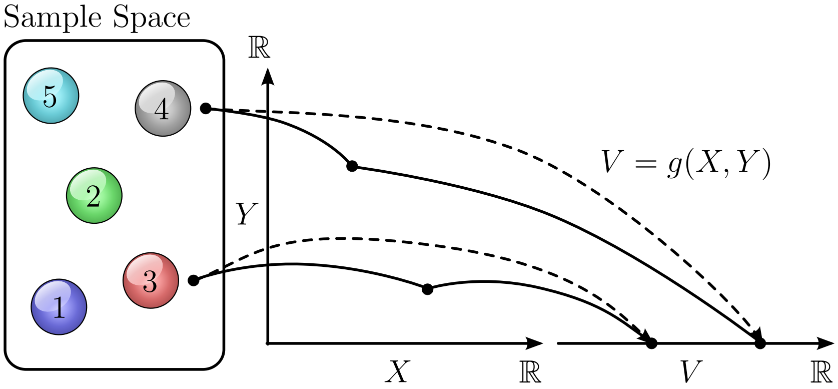

Figure 7.2: A real-valued function of two random variables, X and Y, is also a random variable. Above, $V = g(X, Y)$ maps elements of the sample space to real numbers.

Example 59. *Two dice, a blue die and a red one, are rolled simultaneously. The random variable X represents the number of dots that appears on the top face of the blue die, whereas Y denotes the number of dots on the red die. We can form a random variable U that describes the sum of these two dice, $U = X + Y$.*

The lowest possible value for U is two, and its maximum value is twelve. The PMF of U, as calculated using (7.1), *appears in table form below*

k	$2, 12$	$3, 11$	$4, 10$	$5, 9$	$6, 8$	7
$p_U(k)$	$1/36$	$1/18$	$1/12$	$1/9$	$5/36$	$1/6$

The definition of the expectation operator can be extended to multiple random variables. In particular, the expected value of $g(X, Y)$ is obtained by computing

$$\mathrm{E}[g(X,Y)] = \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} g(x,y) p_{X,Y}(x,y). \tag{7.2}$$

Example 60. *An urn contains three balls numbered one, two and three. Two balls are selected from the urn at random, without replacement. We employ X and Y to represent the numbers on the first and second balls, respectively. We wish to compute the expected value of the function $X + Y$.*

Using (7.2)*, we compute the expectation of $g(X, Y) = X + Y$ as*

$$\begin{aligned} \mathrm{E}[g(X, Y)] &= \mathrm{E}[X + Y] \\ &= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} (x + y) p_{X,Y}(x, y) \\ &= \sum_{x \in X(\Omega)} x p_X(x) + \sum_{y \in Y(\Omega)} y p_Y(y) = 4. \end{aligned}$$

The expected value of $X + Y$ is four.

7.3 Conditional Random Variables

Many events of practical interest are dependent. That is, knowledge about event A may provide partial information about the realization of event B. This inter-dependence is captured by the concept of conditioning, which was first discussed in Chapter 4. In this section, we extend the concept of conditioning to multiple random variables. We study the probability of events concerning random variable Y given that some information about random variable X is available.

Let X and Y be two random variables associated with a same experiment. The *conditional probability mass function* of Y given $X = x$, which we write $p_{Y|X}(\cdot|\cdot)$, is defined by

$$\begin{aligned} p_{Y|X}(y|x) &= \Pr(Y = y | X = x) \\ &= \frac{\Pr(\{Y = y\} \cap \{X = x\})}{\Pr(X = x)} \\ &= \frac{p_{X,Y}(x, y)}{p_X(x)}, \end{aligned}$$

provided that $p_X(x) \neq 0$. Note that conditioning on $X = x$ is not possible when $p_X(x)$ vanishes, as it has no relevant meaning. This is similar to conditional probabilities being only defined for conditional events with non-zero probabilities.

Let x be fixed with $p_X(x) > 0$. The conditional PMF introduced above is a valid PMF since it is nonnegative and

$$\begin{aligned}\sum_{y \in Y(\Omega)} p_{Y|X}(y|x) &= \sum_{y \in Y(\Omega)} \frac{p_{X,Y}(x,y)}{p_X(x)} \\ &= \frac{1}{p_X(x)} \sum_{y \in Y(\Omega)} p_{X,Y}(x,y) = 1.\end{aligned}$$

The probability that Y belongs to S, given $X = x$, is obtained by summing the conditional probability $p_{Y|X}(\cdot|\cdot)$ over all outcomes included in S,

$$\Pr(Y \in S | X = x) = \sum_{y \in S} p_{Y|X}(y|x).$$

Example 61 (Hypergeometric Random Variable). *A wireless communication channel is employed to transmit a data packet that contains a total of m bits. The first n bits of this message are dedicated to the packet header, which stores very sensitive information. The wireless connection is unreliable; every transmitted bit is received properly with probability $1-p$ and erased with probability p, independently of other bits. We wish to derive the conditional probability distribution of the number of erasures located in the header, given that the total number of corrupted bits in the packet is equal to c.*

Let H represent the number of erasures in the header, and denote the number of corrupted bits in the entire message by C. The conditional probability mass function of H given $C = c$ is equal to

$$\begin{aligned}p_{H|C}(h|c) &= \frac{p_{H,C}(h,c)}{p_C(c)} \\ &= \frac{\binom{n}{h}(1-p)^{n-h}p^h \binom{m-n}{c-h}(1-p)^{(m-n)-(c-h)}p^{c-h}}{\binom{m}{c}(1-p)^{m-c}p^c} \\ &= \frac{\binom{n}{h}\binom{m-n}{c-h}}{\binom{m}{c}},\end{aligned}$$

where $h = 0, 1, \ldots \min\{c, n\}$. Clearly, the conditioning affects the probability distribution of the number of corrupted bits in the header. In general, a random variable with such a distribution is known as a hypergeometric random variable.

The definition of conditional PMF can be rearranged to obtain a convenient formula to calculate the joint distribution of X and Y, namely

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x) = p_{X|Y}(x|y)p_Y(y).$$

This formula can be use to compute the joint PMF of X and Y sequentially.

Example 62 (Splitting Property of Poisson PMF)**.** *A digital communication system sends out either a one with probability p or a zero with probability $1-p$, independently of previous transmissions. The number of transmitted binary digits within a given time interval has a Poisson PMF with parameter λ. We wish to show that the number of ones sent in that same time interval has a Poisson PMF with parameter $p\lambda$.*

Let M denote the number of ones within the stipulated interval, N be the number of zeros, and $K = M + N$ be the total number of bits sent during the same interval. The number of ones given that the total number of transmissions is k is given by

$$p_{M|K}(m|k) = \binom{k}{m} p^m (1-p)^{k-m}, \quad m = 0, 1, \ldots, k.$$

The probability that M is equal to m is therefore equal to

$$\begin{aligned} p_M(m) &= \sum_{k=0}^{\infty} p_{K,M}(k,m) = \sum_{k=0}^{\infty} p_{M|K}(m|k) p_K(k) \\ &= \sum_{k=m}^{\infty} \binom{k}{m} p^m (1-p)^{k-m} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{u=0}^{\infty} \binom{u+m}{m} p^m (1-p)^u \frac{\lambda^{u+m}}{(u+m)!} e^{-\lambda} \\ &= \frac{(\lambda p)^m}{m!} e^{-\lambda} \sum_{u=0}^{\infty} \frac{((1-p)\lambda)^u}{u!} = \frac{(\lambda p)^m}{m!} e^{-p\lambda}. \end{aligned}$$

Above, we have used the change of variables $k = u + m$. We have also rearranged the sum into a familiar form, leveraging the fact that the summation of a Poisson PMF over all possible values is equal to one. We can see that M has a Poisson PMF with parameter $p\lambda$.

7.3.1 Conditioning on Events

It is also possible to define the conditional PMF of a random variable X, conditioned on an event S where $\Pr(X \in S) > 0$. Let X be a random variable associated with a particular experiment, and let S be a non-trivial event corresponding to this experiment. The conditional PMF of X given S is defined by

$$p_{X|S}(x) = \Pr(X = x|S) = \frac{\Pr(\{X = x\} \cap S)}{\Pr(S)}. \tag{7.3}$$

Note that the events $\{\omega \in \Omega | X(\omega) = x\}$ form a partition of Ω as x ranges over all the values in $X(\Omega)$. Using the total probability theorem, we gather that

$$\sum_{x \in X(\Omega)} \Pr(\{X = x\} \cap S) = \Pr(S)$$

and, consequently, we get

$$\begin{aligned} \sum_{x \in X(\Omega)} p_{X|S}(x) &= \sum_{x \in X(\Omega)} \frac{\Pr(\{X = x\} \cap S)}{\Pr(S)} \\ &= \frac{\sum_{x \in X(\Omega)} \Pr(\{X = x\} \cap S)}{\Pr(S)} = 1. \end{aligned}$$

We conclude that $p_{X|S}(\cdot)$ is a valid PMF.

Another interpretation of (7.3) is the following. Let $Y = \mathbf{1}_S(\cdot)$ symbolize the indicator function of S, with

$$\mathbf{1}_S(\omega) = \begin{cases} 1, & \omega \in S \\ 0, & \omega \notin S. \end{cases}$$

Then $p_{X|S}(x) = p_{X|Y}(x|1)$. In this sense, conditioning on an event is a special case of a conditional probability mass function.

Example 63. *A data packet is sent to a destination over an unreliable wireless communication link. Data is successfully decoded at the receiver with probability p, and the transmission fails with probability $(1 - p)$, independently of previous trials. When initial decoding fails, a retransmission is requested immediately. However, if packet transmission fails n consecutive times, then the data is dropped from the queue altogether, with no further transmission attempt. Let X denote the number of trials, and let S be the event that the data* *packet is successfully transmitted. We wish to compute the conditional PMF of X given S.*

First, we note that a successful transmission can happen at any of n possible instants. These outcomes being mutually exclusive, the probability of S can be written as

$$\Pr(S) = \sum_{k=1}^{n} (1-p)^{k-1} p = 1 - (1-p)^n .$$

Thus, the conditional PMF $p_{N|S}(\cdot)$ is equal to

$$p_{N|S}(k) = \frac{(1-p)^{k-1} p}{1-(1-p)^n},$$

where $k = 1, 2, \ldots n$.

7.4 Conditional Expectations

The *conditional expectation* of Y given $X = x$ is simply the expectation of Y with respect to the conditional PMF $p_{Y|X}(\cdot|x)$,

$$\mathrm{E}[Y|X=x] = \sum_{y \in Y(\Omega)} y p_{Y|X}(y|x).$$

This conditional expectation can be viewed as a function of x,

$$h(x) = \mathrm{E}[Y|X=x].$$

It is therefore mathematically accurate and sometimes desirable to talk about the random variable $h(X) = \mathrm{E}[Y|X]$. In particular, if X and Y are two random variables associated with an experiment, the outcome of this experiment determines the value of X, say $X = x$, which in turn yields the conditional expectation $h(x) = \mathrm{E}[Y|X=x]$. From this point of view, the conditional expectation $\mathrm{E}[Y|X]$ is simply an instance of a random variable.

Not too surprisingly, the expectation of $\mathrm{E}[Y|X]$ is equal to $\mathrm{E}[Y]$, as evinced by the following derivation,

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}[Y|X]\right] &= \sum_{x \in X(\Omega)} \mathrm{E}[Y|X=x] p_X(x) \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} y p_{Y|X}(y|x) p_X(x) \\
&= \sum_{y \in Y(\Omega)} \sum_{x \in X(\Omega)} y p_{X,Y}(x,y) \\
&= \sum_{y \in Y(\Omega)} y p_Y(y) = \mathrm{E}[Y].
\end{aligned}$$

Using a similar argument, it is straightforward to show that

$$\mathrm{E}\left[\mathrm{E}[g(Y)|X]\right] = \mathrm{E}[g(Y)].$$

Example 64. *An entrepreneur opens a small business that sells two kinds of beverages from the Brazos Soda Company, cherry soda and lemonade. The number of bottles sold in an hour at the store is found to be a Poisson random variable with parameter $\lambda = 10$. Every customer selects a cherry soda with probability p and a lemonade with probability $(1-p)$, independently of other customers. We wish to find the conditional mean of the number of cherry sodas purchased in an hour given that ten beverages were sold to customers during this time period.*

Let B represent the number of bottles sold during an hour. Similarly, let C and L be the number of cherry sodas and lemonades purchased during the same time interval, respectively. We note that $B = C + L$. The conditional PMF of C given that the total number of beverages sold equals ten is

$$p_{C|B}(k|10) = \binom{10}{k} p^k (1-p)^{10-k}. \tag{7.4}$$

This follows from the fact that every customer selects a cherry soda with probability p, independently of other customers. The conditional mean is then seen to equal

$$\mathrm{E}[C|B=10] = \sum_{k=0}^{10} k \binom{10}{k} p^k (1-p)^{10-k} = 10p.$$

We can define the expectation of X conditioned on event S in an analogous fashion. Let S be an event such that $\Pr(S) > 0$. The conditional expectation of X given S is

$$\mathrm{E}[X|S] = \sum_{x \in X(\Omega)} x p_{X|S}(x),$$

where $p_{X|S}(\cdot)$ is defined in (7.3). Similarly, we have

$$\mathrm{E}[g(X)|S] = \sum_{x \in X(\Omega)} g(x) p_{X|S}(x),$$

Example 65. *Spring break is coming and a student decides to renew his shirt collection. The number of shirts purchased by the student is a random variable denoted by N. The PMF of this random variable is a geometric distribution with parameter $p = 0.5$. Any one shirt costs \$10, \$20 or \$50 with respective probabilities 0.5, 0.3 and 0.2, independently of other shirts. We wish to compute the expected amount of money spent by the student during his shopping spree. Also, we wish to compute the expected amount of money disbursed given that the student buys at least five shirts.*

Let C_i be the cost of the ith shirt. The total amount of money spent by the student, denoted by T, can be expressed as

$$T = \sum_{i=1}^{N} C_i.$$

The mean of T can be computed using nested conditional expectation. It is equal to

$$\begin{aligned}
\mathrm{E}[T] &= \mathrm{E}\left[\sum_{i=1}^{N} C_i\right] = \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{N} C_i \middle| N\right]\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{N} \mathrm{E}[C_i|N]\right] = 21\mathrm{E}[N] = 42.
\end{aligned}$$

The student is expected to spend \$42. Given that the student buys at least five shirts, the conditional expectation becomes

$$\begin{aligned}
\mathrm{E}[T|N \geq 5] &= \mathrm{E}\left[\sum_{i=1}^{N} C_i \middle| N \geq 5\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{N} C_i \middle| N\right] \middle| N \geq 5\right] \\
&= 21\mathrm{E}[N|N \geq 5] = 126.
\end{aligned}$$

Conditioned on buying more than five shirts, the student is expected to spend \$126. Note that, in computing the conditional expectation, we have utilized the memoryless property of the geometric random variable,

$$\mathrm{E}[N|N \geq 5] = \mathrm{E}[N|N > 4] = 4 + \mathrm{E}[N] = 6.$$

7.5 Independence

Let X and Y be two random variables associated with a same experiment. We say that X and Y are *independent random variables* if

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

for every $x \in X(\Omega)$ and every $y \in Y(\Omega)$.

There is a clear relation between the concept of independence introduced in Section 4.4 and the independence of two random variables. Random variables X and Y are independent if and only if the events $\{X = x\}$ and $\{Y = y\}$ are independent for every pair (x, y) such that $x \in X(\Omega)$ and $y \in Y(\Omega)$.

Proposition 6. *If X and Y are independent random variables, then*

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

Proof. Assume that both $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ exist, then

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xy p_{X,Y}(x, y) \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xy p_X(x) p_Y(y) \\
&= \sum_{x \in X(\Omega)} x p_X(x) \sum_{y \in Y(\Omega)} y p_Y(y) = \mathrm{E}[X]\mathrm{E}[Y],
\end{aligned}$$

where we have used the fact that $p_{X,Y}(x, y) = p_X(x) p_Y(y)$ for independent random variables. □

Example 66. *Two dice of different colors are rolled, as in Example 59. We wish to compute the expected value of their product. We know that the mean of each role is* $\mathrm{E}[X] = \mathrm{E}[Y] = 3.5$. *Furthermore, it is straightforward to show* *that X and Y, the numbers of dots on the two dice, are independent random variables. The expected value of their product is then equal to the product of the individual means,* $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y] = 12.25$.

We can parallel the proof of Proposition 6 to show that

$$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\mathrm{E}[h(Y)] \tag{7.5}$$

whenever X and Y are independent random variables and the corresponding expectations exist.

7.5.1 Sums of Independent Random Variables

We turn to the question of determining the distribution of a sum of two independent random variables in terms of the marginal PMF of the summands. Suppose X and Y are independent random variables that take on integer values. Let $p_X(\cdot)$ and $p_Y(\cdot)$ be the PMFs of X and Y, respectively. We wish to determine the distribution $p_U(\cdot)$ of U, where $U = X + Y$. To accomplish this task, it suffices to compute the probability that U assumes a value k, where k is an arbitrary integer,

$$\begin{aligned} p_U(k) &= \Pr(U = k) = \Pr(X + Y = k) \\ &= \sum_{\{(x,y) \in X(\Omega) \times Y(\Omega) | x+y=k\}} p_{X,Y}(x, y) \\ &= \sum_{m \in \mathbb{Z}} p_{X,Y}(m, k - m) \\ &= \sum_{m \in \mathbb{Z}} p_X(m) p_Y(k - m). \end{aligned}$$

The latter operation is called a *discrete convolution*. In particular, the PMF of $U = X + Y$ is the discrete convolution of $p_X(\cdot)$ and $p_Y(\cdot)$ given by

$$p_U(k) = (p_X * p_Y)(k) = \sum_{m \in \mathbb{Z}} p_X(m) p_Y(k - m). \tag{7.6}$$

The discrete convolution is commutative and associative.

One interesting application of (7.5) occurs when dealing with sums of independent random variables. Suppose X and Y are two independent random variables that take on integer values and, again, let $U = X + Y$. The ordinary generating function of U, as defined in Section 6.4, is given by

$$\begin{aligned} G_U(z) &= \mathrm{E}\left[z^U\right] = \mathrm{E}\left[z^{X+Y}\right] \\ &= \mathrm{E}\left[z^X z^Y\right] = \mathrm{E}\left[z^X\right]\mathrm{E}\left[z^Y\right] \\ &= G_X(z)G_Y(z). \end{aligned}$$

That is, the generating function of a sum of independent random variables is equal to the product of the individual ordinary generating functions.

Example 67 (Sum of Poisson Random Variables)**.** *Let X be a Poisson random variable with parameter α and let Y be a Poisson random variable with parameter β. We wish to find the probability mass function of $U = X + Y$.*

To solve this problem, we use ordinary generating functions. First, recall that the generating function of a Poisson random variable with parameter λ is $e^{\lambda(z-1)}$. The ordinary generating functions of X and Y are therefore equal to

$$\begin{aligned} G_X(z) &= e^{\alpha(z-1)} \\ G_Y(z) &= e^{\beta(z-1)}. \end{aligned}$$

As such, the ordinary generating function of $U = X + Y$ is

$$G_U(z) = G_X(z)G_Y(z) = e^{\alpha(z-1)}e^{\beta(z-1)} = e^{(\alpha+\beta)(z-1)}.$$

We conclude, by the uniqueness of generating functions, that U is a Poisson random variable with parameter $\alpha + \beta$ and, accordingly, we get

$$p_U(k) = \frac{(\alpha+\beta)^k}{k!}e^{-(\alpha+\beta)}, \quad k = 0, 1, 2, \ldots$$

This method of finding the PMF of U is more concise than using the discrete convolution.

We saw in Section 7.3 that a random variable can also be conditioned on a specific event. Let X be a random variable and let S be a non-trivial event. The variable X is *independent* of S if

$$\Pr(\{X = x\} \cap S) = p_X(x)\Pr(S)$$

for every $x \in X(\Omega)$. In particular, if X is independent of event S then

$$p_{X|S}(x) = p_X(x)$$

for all $x \in X(\Omega)$.

7.6 Numerous Random Variables

The notion of a joint distribution can be applied to any number of random variables. Let $X_1, X_2, \ldots, X_n$ be random variables; their joint PMF is defined by

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \Pr\left(\bigcap_{k=1}^{n} \{X_k = x_k\} \right),$$

or, in vector form, $p_{\mathbf{X}}(\mathbf{x}) = \Pr(\mathbf{X} = \mathbf{x})$. When the random variables $\{X_k\}$ are independent, their joint PMF reduces to

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^{n} p_{X_k}(x_k).$$

Although not discussed in details herein, we emphasize that most concepts introduced in this chapter can be extended to multiple random variables in a straightforward manner. Also, we note that, from an abstract perspective, the space defined by vectors of the form $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$, $\omega \in \Omega$, is itself a sample space on which random variables can be defined.

Consider the empirical sums

$$S_n = \sum_{k=1}^{n} X_k \quad n \geq 1,$$

where $X_1, X_2, \ldots$ are independent integer-valued random variables, each with marginal PMF $p_X(\cdot)$. Obviously, the distribution of S_1 is simply $p_X(\cdot)$. More generally, the PMF of S_n can be obtained recursively using the formula

$$S_n = S_{n-1} + X_n.$$

This leads to the PMF

$$p_{S_n}(k) = (p_X * p_X * \cdots * p_X)(k),$$

which is the n-fold convolution of $p_X(\cdot)$. Furthermore, the ordinary generating function of S_n can be written as

$$G_{S_n}(z) = (G_X(z))^n.$$

Example 68 (Binomial Random Variables)**.** *Suppose that S_n is a sum of n independent and identically distributed Bernoulli random variables, each with parameter $p \in (0, 1)$. The PMF of S_1 is equal to*

$$p_{S_1}(k) = \begin{cases} 1 - p & k = 0 \\ p & k = 1. \end{cases}$$

Assume that the PMF of S_{n-1} is given by

$$p_{S_{n-1}}(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}, \quad k = 0, 1, \ldots n-1.$$

Then, the distribution of S_n can be computed recursively using the discrete convolution,

$$\begin{aligned} p_{S_n}(k) &= \sum_{m=-\infty}^{\infty} p_{S_{n-1}}(m) p_X(k-m) \\ &= p_{S_{n-1}}(k)(1-p) + p_{S_{n-1}}(k-1)p \\ &= \binom{n-1}{k} p^k (1-p)^{n-k} + \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

where $k = 0, 1, \ldots, n$. Thus, by the principle of mathematical induction, we gather that the sum of n independent Bernoulli random variables, each with parameter p, is a binomial random variable with parameters n and p.

The convolution of two binomial distributions, one with parameter m and p and the other with parameters n and p, is also a binomial random distribution with parameters $(m + n)$ and p. This fact follows directly from the previous argument.

Example 69 (Negative Binomial Random Variable*)**.** *Suppose that a Bernoulli trial is repeated multiple times until r ones are obtained. We denote by X the random variable that represents the number of zeros observed before completion of the process. The distribution of X is given by*

$$p_X(k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \quad k = 0, 1, \ldots$$

where p is the common parameter of the Bernoulli trials. In general, a random variable with such a distribution is known as a negative binomial random variable.

We wish to show that X can be obtained as a sum of r independent random variables, a task which we perform by looking at its ordinary generating function. We use properties of differential equations to derive $G_X(z)$ from $p_X(\cdot)$. By definition, we have

$$G_X(z) = \sum_{k=0}^{\infty} z^k \binom{k+r-1}{r-1} p^r (1-p)^k.$$

Consider the derivative of $G_X(z)$,

$$\begin{aligned}
\frac{dG_X}{dz}(z) &= \sum_{k=0}^{\infty} k z^{k-1} \binom{k+r-1}{r-1} p^r (1-p)^k \\
&= \sum_{k=1}^{\infty} z^{k-1} \frac{(k+r-1)!}{(r-1)!(k-1)!} p^r (1-p)^k \\
&= (1-p) \sum_{\ell=0}^{\infty} (\ell + r) z^{\ell} \binom{\ell+r-1}{r-1} p^r (1-p)^{\ell} \\
&= (1-p) \left(z \frac{dG_X}{dz}(z) + r G_X(z) \right).
\end{aligned}$$

It follows from this equation that

$$\frac{1}{G_X(z)} \frac{dG_X}{dz}(z) = \frac{(1-p)r}{1-(1-p)z}$$

or, alternatively, we can write

$$\frac{d}{dz} \log \left(G_X(z) \right) = \frac{(1-p)r}{1-(1-p)z}.$$

Integrating both sides yields

$$\log \left(G_X(z) \right) = -r \log \left(1 - (1-p)z \right) + c,$$

where c is an arbitrary constant. Applying boundary condition $G_X(1) = 1$, we get the ordinary generating function of X as

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

From this equation, we can deduce that X is the sum of r independent random variables, $Y_1, \ldots, Y_r$, each with ordinary generating function

$$G_{Y_i}(z) = \frac{p}{1-(1-p)z}.$$

In particular, the distribution of Y_i is given by

$$p_{Y_i}(m) = \frac{1}{m!} \frac{d^m G_{Y_i}}{dz^m}(0) = p(1-p)^m \quad m = 0, 1, \ldots$$

It may be instructive to compare this distribution with the PMF of a geometric random variable.

Further Reading

1. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Sections 2.5–2.7.

2. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Sections 6.1–6.4.

3. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 3.4–3.5.

Chapter 8

Continuous Random Variables

So far, we have studied discrete random variables and we have explored their properties. Discrete random variables are quite useful in many contexts, yet they form only a small subset of the collection of random variables pertinent to applied probability and engineering. In this chapter, we consider random variables that range over a continuum of possible values; that is, random variables that can take on an uncountable set of values.

Continuous random variables are powerful mathematical abstractions that allow engineers to pose and solve important problems. Many of these problems are difficult to address using discrete models. While this extra flexibility is useful and desirable, it comes at a certain cost. A continuous random variable cannot be characterized by a probability mass function. This predicament emerges from the limitations of the third axiom of probability laws, which only applies to countable collections of disjoint events.

Below, we provide a definition for continuous random variables. Furthermore, we extend and apply the concepts and methods initially developed for discrete random variables to the class of continuous random variables. In particular, we develop a continuous counterpart to the probability mass function.

8.1 Cumulative Distribution Functions

We begin our exposition of continuous random variables by introducing a general concept which can be employed to bridge our understanding of discrete and continuous random variables. Recall that a *random variable* is a real-valued function acting on the outcomes of an experiment. In particular, given a sample space, random variable X is a function from Ω to $\mathbb{R}$. The *cumulative distribution function (CDF)* of X is defined point-wise as the probability of the event $\{X \leq x\}$,

$$F_X(x) = \Pr(\{X \leq x\}) = \Pr(X \leq x).$$

In terms of the underlying sample space, $F_X(x)$ denotes the probability of the set of all outcomes in Ω for which the value of X is less than or equal to x,

$$F_X(x) = \Pr\left(X^{-1}((-\infty, x])\right) = \Pr(\{\omega \in \Omega | X(\omega) \leq x\}).$$

In essence, the CDF is a convenient way to specify the probability of all events of the form $\{X \in (-\infty, x]\}$.

The CDF of random variable X exists for any well-behaved function $X : \Omega \mapsto \mathbb{R}$. Moreover, since the realization of X is a real number, we have

$$\begin{aligned} \lim_{x \downarrow -\infty} F_X(x) &= 0 \\ \lim_{x \uparrow \infty} F_X(x) &= 1. \end{aligned}$$

Suppose $x_1 < x_2$, then we can write $\{X \leq x_2\}$ as the union of the two disjoint sets $\{X \leq x_1\}$ and $\{x_1 < X \leq x_2\}$. It follows that

$$\begin{aligned} F_X(x_2) &= \Pr(X \leq x_2) \\ &= \Pr(X \leq x_1) + \Pr(x_1 < X \leq x_2) \\ &\geq \Pr(X \leq x_1) = F_X(x_1). \end{aligned} \tag{8.1}$$

In other words, a CDF is always a non-decreasing function. Finally, we note from (8.1) that the probability of X falling in the interval $(x_1, x_2]$ is

$$\Pr(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1). \tag{8.2}$$

8.1.1 Discrete Random Variables

If X is a discrete random variable, then the CDF of X is given by

$$F_X(x) = \sum_{u \in X(\Omega) \cap (-\infty, x]} p_X(u),$$

and its PMF can be computed using the formula

$$p_X(x) = \Pr(X \leq x) - \Pr(X < x) = F_X(x) - \lim_{u \uparrow x} F_X(u).$$

Fortunately, this formula is simpler when the random variable X only takes integer values, as seen in the example below.

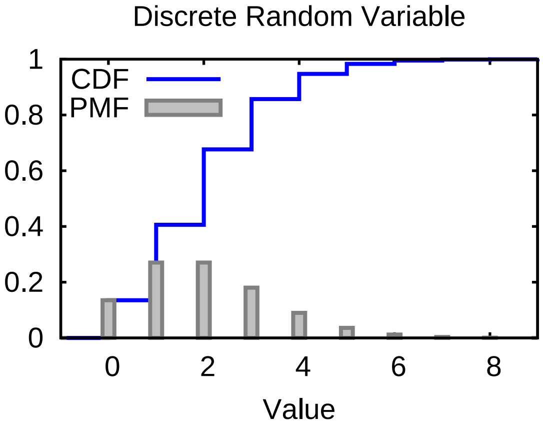

Figure 8.1: This figure shows the PMF of a discrete random variable, along with the corresponding CDF. The values of the PMF are depicted by the height of the rectangles; their cumulative sums lead to the values of the CDF.

Example 70. *Let X be a geometric random variable with parameter p,*

$$p_X(k) = (1-p)^{k-1}p \quad k = 1, 2, \ldots$$

For $x > 0$, the CDF of X is given by

$$F_X(x) = \sum_{k=1}^{\lfloor x \rfloor} (1-p)^{k-1}p = 1 - (1-p)^{\lfloor x \rfloor},$$

where $\lfloor \cdot \rfloor$ denotes the standard floor function. For integer $k \geq 1$, the PMF of geometric random variable X can be recovered from the CDF as follows,

$$\begin{aligned} p_X(k) &= F_X(x) - \lim_{u \uparrow x} F_X(u) = F_X(k) - F_X(k-1) \\ &= \left(1 - (1-p)^k\right) - \left(1 - (1-p)^{k-1}\right) \\ &= (1-p)^{k-1}p. \end{aligned}$$

8.1.2 Continuous Random Variables

Having introduced the general notion of a CDF, we can safely provide a more precise definition for continuous random variables. Let X be a random variable with CDF $F_X(\cdot)$, then X is said to be a *continuous random variable* if $F_X(\cdot)$ is continuous and differentiable.

Example 71. *Suppose that X is a random variable with CDF given by*

$$F_X(x) = \begin{cases} 1 - e^{-x}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

This cumulative distribution function is differentiable with

$$\frac{dF_X}{dx}(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & x < 0 \end{cases}$$

and therefore X is a continuous random variable.

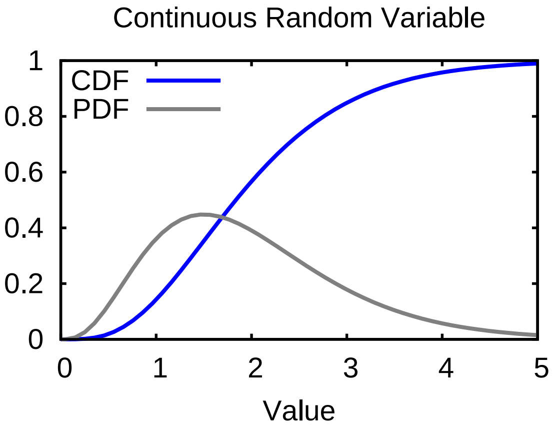

Figure 8.2: The CDF of a continuous random variable is differentiable. This figure provides one example of a continuous random variable. Both, the CDF $F_X(\cdot)$ and its derivative $f_X(\cdot)$ (PDF) are displayed.

8.1.3 Mixed Random Variables*

Generally speaking, the CDF of a discrete random variable is a discontinuous staircase-like function, whereas the CDF of a continuous random variable is continuous and differentiable almost everywhere. There exist random variables for which neither situation applies. Such random variables are sometimes called *mixed random variables*. Our exposition of mixed random variables in this document is very limited. Still, we emphasize that a good understanding of discrete and continuous random variables is instrumental in understanding and solving problems including mixed random variables.

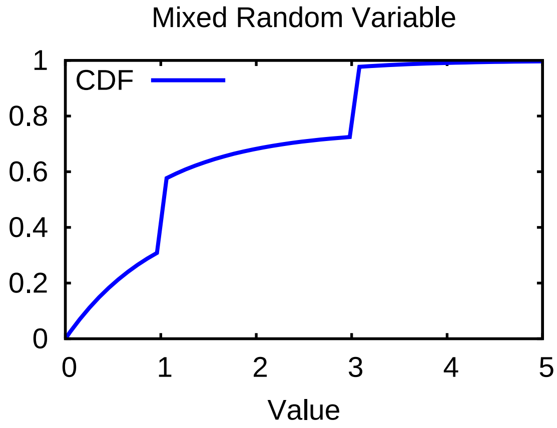

Figure 8.3: This figure shows the CDF of a mixed random variable. In general, mixed random variables do not have a PMF nor a PDF. Their CDF may be composed of a mixture of differentiable intervals and discontinuous jumps.

8.2 Probability Density Functions

As mentioned above, the CDF of continuous random variable X is a differentiable function. The derivative of $F_X(\cdot)$ is called the *probability density function (PDF)* of X, and it is denoted by $f_X(\cdot)$. If X is a random variable with PDF $f_X(\cdot)$ then, by the fundamental theorem of calculus, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi) d\xi.$$

Equivalently, we can write

$$f_X(x) = \frac{dF_X}{dx}(x).$$

Note that PDFs are only defined for continuous random variables. This is somewhat restrictive. Nevertheless, the PDF can be a very powerful tool to derive properties of continuous random variables, which may otherwise be difficult to compute.

For $x_1 < x_2$, we can combine the definition of $f_X(\cdot)$ and (8.2) to obtain

$$\Pr(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(\xi) d\xi.$$

Furthermore, it is easily seen that for any continuous random variable

$$\begin{aligned}\Pr(X = x_2) &= \lim_{x_1 \uparrow x_2} \Pr(x_1 < X \leq x_2) = \lim_{x_1 \uparrow x_2} \int_{x_1}^{x_2} f_X(\xi) d\xi \\ &= \int_{x_2}^{x_2} f_X(\xi) d\xi = 0.\end{aligned}$$

In other words, if X is a continuous random variable, then $\Pr(X = x) = 0$ for any real number x. An immediate corollary of this fact is

$$\begin{aligned}\Pr(x_1 < X < x_2) &= \Pr(x_1 \leq X < x_2) \\ &= \Pr(x_1 < X \leq x_2) = \Pr(x_1 \leq X \leq x_2);\end{aligned}$$

the inclusion or exclusion of endpoints in an interval does not affect the probability of the corresponding interval when X is a continuous random variable.

We can derive properties for the PDF of continuous random variable X based on the axioms of probability laws. First, the probability that X is a real number is given by

$$\Pr(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(\xi) d\xi = 1.$$

Thus, $f_X(x)$ must integrate to one. Also, because the probabilities of events are nonnegative, we must have $f_X(x) \geq 0$ everywhere. Finally, given an admissible set S, the probability that $X \in S$ can be expressed through an integral,

$$\Pr(X \in S) = \int_S f_X(\xi) d\xi.$$

Admissible events of the form $\{X \in S\}$ are sets for which we know how to carry this integral.

8.3 Important Distributions

Good intuition about continuous random variables can be developed by looking at examples. In this section, we introduce important random variables and their distributions. These random variables find widespread application in various fields of engineering.

8.3.1 The Uniform Distribution

A (continuous) *uniform random variable* is such that all intervals of a same length contained within its support are equally probable. The PDF of a uniform random variable is defined by two parameters, a and b, which represent the minimum and maximum values of its support, respectively. The PDF $f_X(\cdot)$ is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise.} \end{cases}$$

The associated cumulative distribution function becomes

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x \geq b. \end{cases}$$

Example 72. *David comes to campus every morning riding the Aggie Spirit Transit. On his route, a bus comes every thirty minutes, from sunrise until dusk. David, who does not believe in alarm clocks or watches, wakes up with daylight. After cooking a hefty breakfast, he walks to the bus stop. If his arrival time at the bus stop is uniformly distributed between 9:00 a.m. and 9:30 a.m., what is the probability that he waits less than five minutes for the bus?*

Let t_0 be the time at which David arrives at the bus stop, and denote by T the time he spends waiting. The time at which the next bus arrives at David's stop is uniformly distributed between t_0 and $t_0 + 30$. The amount of time that he spends at the bus stop is therefore uniformly distributed between 0 and 30 minutes. Accordingly, we have

$$f_T(t) = \begin{cases} \frac{1}{30}, & t \in [0, 30] \\ 0, & \textit{otherwise.} \end{cases}$$

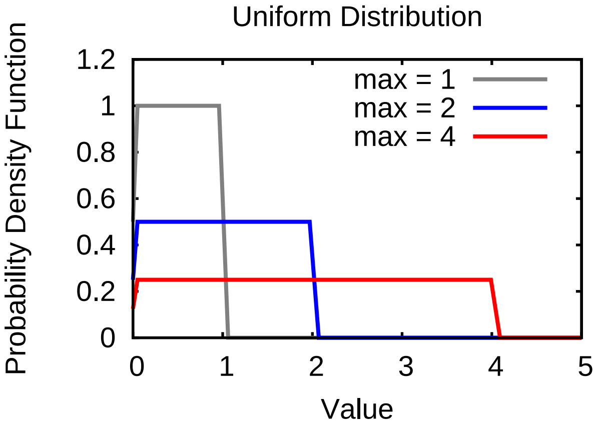

Figure 8.4: This figure shows the PDFs of uniform random variables with support intervals $[0,1]$, $[0,2]$ and $[0,4]$.

The probability that David waits less than five minutes is

$$\Pr(T<5)=\int_0^5 \frac{1}{30}dt=\frac{1}{6}.$$

8.3.2 The Gaussian (Normal) Random Variable

The *Gaussian random variable* is of fundamental importance in probability and statistics. It is often used to model distributions of quantities influenced by large numbers of small random components. The PDF of a Gaussian random variable is given by

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad -\infty<x<\infty,$$

where m and $\sigma>0$ are real parameters.

A Gaussian variable whose distribution has parameters $m=0$ and $\sigma=1$ is called a *normal random variable* or a *standard Gaussian random variable*, names that hint at its popularity. The CDF of a Gaussian random variable does not admit a closed-form expression; it can be expressed as

$$\begin{aligned}F_X(x)=\Pr(X\leq x)&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(\xi-m)^2}{2\sigma^2}}d\xi\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(x-m)/\sigma} e^{-\frac{\zeta^2}{2}}d\zeta=\Phi\left(\frac{x-m}{\sigma}\right),\end{aligned}$$

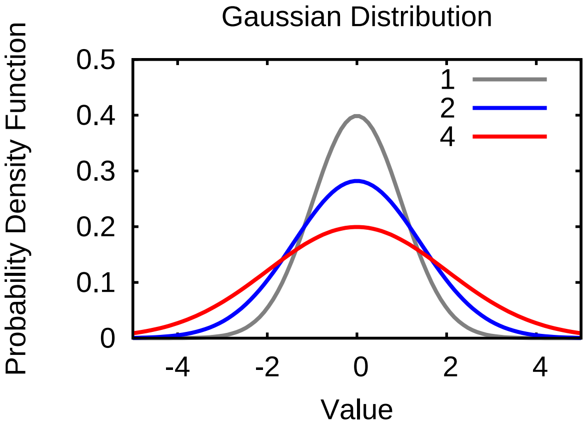

Figure 8.5: The distributions of Gaussian random variables appear above for parameters $m = 0$ and $\sigma^2 \in \{1, 2, 4\}$.

where $\Phi(\cdot)$ is termed the *standard normal cumulative distribution function* and is defined by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\zeta^2}{2}} d\zeta.$$

We emphasize that the function $\Phi(\cdot)$ is nothing more than a convenient notation for the CDF of a normal random variable.

Example 73. *A binary signal is transmitted through a noisy communication channel. The sent signal takes on either a value of* 1 *or* -1. *The message received at the output of the communication channel is corrupted by additive thermal noise. This noise can be accurately modeled as a Gaussian random variable. The receiver declares that a* 1 *(*-1*) was transmitted if the sent signal is positive (negative). What is the probability of making an erroneous decision?*

Let $S \in \{-1, 1\}$ *denote the transmitted signal,* N *be the value of the thermal noise, and* Y *represent the value of the received signal. An error can occur in one of two possible ways:* $S = 1$ *was transmitted and* Y *is less than zero, or* $S = -1$ *was transmitted and* Y *is greater than zero. Using the total probability theorem, we can compute the probability of error as*

$$\Pr(Y \geq 0 | S = -1) \Pr(S = -1) + \Pr(Y \leq 0 | S = 1) \Pr(S = 1).$$

By symmetry, it is easily argued that

$$\begin{aligned} \Pr(Y \leq 0 | S = 1) &= \Pr(Y \geq 0 | S = -1) = \Pr(N > 1) \\ &= \int_1^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\xi^2}{2\sigma^2}} d\xi = 1 - \Phi\left(\frac{1}{\sigma}\right). \end{aligned}$$

The probability that the receiver makes an erroneous decision is $1 - \Phi(1/\sigma)$. *The reliability of this transmission scheme depends on the amount of noise present at the receiver.*

The normal random variable is so frequent in applied mathematics and engineering that many variations of its CDF possess their own names. The *error function* is a function which is primarily encountered in the fields of statistics and partial differential equations. It is defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2} d\xi.$$

The error function is related to the standard normal cumulative distribution function by scaling and translation,

$$\Phi(x) = \frac{1 + \operatorname{erf}\left(x/\sqrt{2}\right)}{2}.$$

If X is a standard normal random variable, then $\operatorname{erf}\left(x/\sqrt{2}\right)$ denotes the probability that X lies in the interval $(-x, x)$. In engineering, it is customary to employ the *Q-function*, which is given by

$$\begin{aligned} Q(x) &= \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{\xi^2}{2}} d\xi = 1 - \Phi(x) \\ &= \frac{1 - \operatorname{erf}\left(x/\sqrt{2}\right)}{2}. \end{aligned} \tag{8.3}$$

Equation (8.3) may prove useful when using software packages that provide a built-in implementation for $\operatorname{erf}(\cdot)$, but not for the Q-function. The probability of an erroneous decision in Example 73 can be expressed concisely using the Q-function as $Q(1/\sigma)$.

Next, we prove that the standard normal PDF integrates to one. The solution is easy to follow, but hard to discover. It is therefore useful to include it in this document. Consider a standard normal PDF,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

We can show that $f_X(x)$ integrates to one using a subtle argument and a change of variables. We start with the square of the integrated PDF and proceed from there,

$$\begin{aligned}\left(\int_{-\infty}^{\infty} f_X(\xi) d\xi \right)^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} d\xi \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\zeta^2}{2}} d\zeta \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{\xi^2+\zeta^2}{2}} d\xi d\zeta = \int_0^{2\pi} \frac{1}{2\pi} d\theta \int_0^{\infty} e^{-\frac{r^2}{2}} r dr \\ &= \left(-e^{-\frac{r^2}{2}} \right) \Big|_0^{\infty} = 1.\end{aligned}$$

Since the square of the desired integral is nonnegative and equal to one, we can conclude that the normal PDF integrates to one.

8.3.3 The Exponential Distribution

The *exponential random variable* is also frequently encountered in engineering. It can be used to model the lifetime of devices and systems, and the time elapsed between specific occurrences. An exponential random variable X with parameter $\lambda > 0$ has PDF

$$f_X(x) = \lambda e^{-\lambda x} \quad x \geq 0.$$

For $x \geq 0$, its CDF is equal to

$$F_X(x) = 1 - e^{-\lambda x}.$$

The parameter λ characterizes the rate at which events occur.

Example 74. *Connection requests at an Internet server are characterized by an exponential inter-arrival time with parameter $\lambda = 1/2$. If a request arrives at time t_0, what is the probability that the next packet arrives within two minutes?*

The probability that the inter-arrival time T is less than two minutes can be computed as

$$\begin{aligned}\Pr(T < 2) &= \int_0^2 \frac{1}{2} e^{-\frac{\xi}{2}} d\xi = -e^{-\frac{\xi}{2}} \Big|_0^2 \\ &= 1 - e^{-1} \approx 0.632.\end{aligned}$$

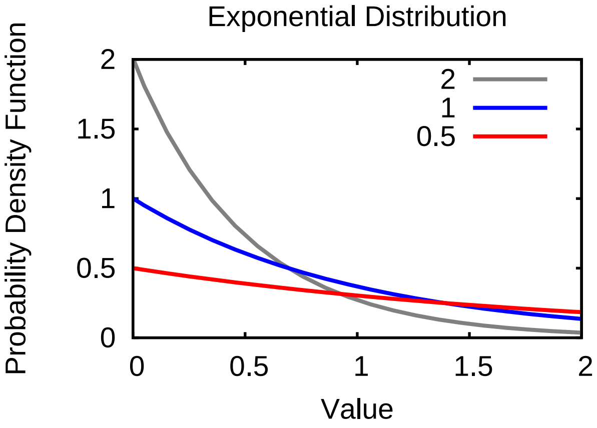

Figure 8.6: The distributions of exponential random variables are shown above for parameters $\lambda \in \{0.5, 1, 2\}$.

The exponential random variable can be obtained as the limit of a sequence of geometric random variables. Let λ be fixed and defined $p_n = \lambda/n$. We define the PMF of random variable Y_n as

$$p_{Y_n}(k) = (1 - p_n)^{k-1} p_n = \left(1 - \frac{\lambda}{n}\right)^{k-1} \frac{\lambda}{n} \quad k = 1, 2, \ldots$$

That is, random variable Y_n is a standard geometric random variable with parameter $p_n = \lambda/n$. For every n, we create a new variable X_n,

$$X_n = \frac{Y_n}{n}.$$

By construction, the random variable X_n has PMF

$$p_{X_n}(y) = \begin{cases} (1 - p_n)^{k-1} p_n, & \text{if } y = k/n \\ 0, & \text{otherwise.} \end{cases}$$

For any $x \geq 0$, the CDF of random variable X_n can be computed as

$$\begin{aligned} \Pr(X_n \leq x) = \Pr(Y_n \leq nx) &= \sum_{k=1}^{\lfloor nx \rfloor} p_{Y_n}(k) \\ &= \sum_{k=1}^{\lfloor nx \rfloor} (1 - p_n)^{k-1} p_n = 1 - (1 - p_n)^{\lfloor nx \rfloor}. \end{aligned}$$

In the limit, as n grows unbounded, we get

$$\begin{aligned}\lim_{n\to\infty} \Pr(X_n \leq x) &= \lim_{n\to\infty} \left[1 - (1-p_n)^{\lfloor nx \rfloor}\right] \\ &= 1 - \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{\lfloor nx \rfloor} \\ &= 1 - e^{-\lambda x}.\end{aligned}$$

Thus, the sequence of scaled geometric random variables $\{X_n\}$ converges in distribution to an exponential random variable X with parameter λ.

Memoryless Property: In view of this asymptotic characterization and the fact that geometric random variables are memoryless, it is not surprising that the exponential random variable also satisfies the *memoryless property*,

$$\Pr(X > t + u | X > t) = \Pr(X > u).$$

This fact can be shown by a straightforward application of conditional probability. Suppose that X is an exponential random variable with parameter λ. Also, let t and u be two positive numbers. The memoryless property can be verified by expanding the conditional probability of X using definition (4.2),

$$\begin{aligned}\Pr(X > t + u | X > t) &= \frac{\Pr(\{X > t + u\} \cap \{X > t\})}{\Pr(X > t)} \\ &= \frac{\Pr(X > t + u)}{\Pr(X > t)} = \frac{e^{-\lambda(t+u)}}{e^{-\lambda t}} \\ &= e^{-\lambda u} = \Pr(X > u).\end{aligned}$$

In reality, the exponential random variable is the only continuous random variable that satisfies the memoryless property.

Example 75. *A prominent company, Century Oak Networks, maintains a bank of servers for its operation. Hard drives on the servers have a half-life of two years. We wish to compute the probability that a specific disk needs repair within its first year of usage.*

Half-lives are typically used to describe quantities that undergo exponential decay. Let T denote the time elapsed until failure of the disk. We know that T is an exponential random variable and, although we are not given λ explicitly, we know that

$$\Pr(T > 2) = \frac{1}{2}.$$

We use the memoryless property to solve this problem,

$$\begin{aligned}\Pr(T > 2) &= \Pr(T > 1)\Pr(T > 1 + 1 | T > 1) \\ &= \Pr(T > 1)\Pr(T > 1) = \left(\Pr(T > 1)\right)^2 .\end{aligned}$$

It follows that $\Pr(T > 1) = \sqrt{\Pr(T > 2)} = 1/\sqrt{2}$*. We can then write* $\Pr(T < 1) = 1 - \Pr(T > 1) \approx 0.293$*. An alternative way to solve this problem would be to first find the value of* λ *associated with* T*, and then compute* $\Pr(T < 1)$ *from the corresponding integral.*

8.4 Additional Distributions

Probability distributions arise in many different contexts and they assume various forms. We conclude this first chapter on continuous random variables by mentioning a few additional distributions that find application in engineering. It is interesting to note the interconnection between various random variables and their corresponding probability distributions.

8.4.1 The Gamma Distribution

The gamma PDF defines a versatile collection of distributions. The PDF of a *gamma random variable* is given by

$$f_X(x) = \frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)} \quad x > 0,$$

where $\Gamma(\cdot)$ denotes the gamma function defined by

$$\Gamma(z) = \int_0^\infty \xi^{z-1}e^{-\xi}d\xi \quad z > 0.$$

The two parameters $\alpha > 0$ and $\lambda > 0$ affect the shape of the ensuing distribution significantly. By varying these two parameters, it is possible for the gamma PDF to accurately model a wide array of empirical data.

The gamma function can be evaluated recursively using integration by parts; this yields the relation $\Gamma(z + 1) = z\Gamma(z)$ for $z > 0$. For nonnegative integers, it can easily be shown that $\Gamma(k+1) = k!$. Perhaps, the most well-known value for the gamma function at a non-integer argument is $\Gamma(1/2) = \sqrt{\pi}$. Interestingly, this specific value for the gamma function can be evaluated by a procedure similar to the one we used to integrate the Gaussian distribution,

$$\begin{aligned}\left(\Gamma\left(\frac{1}{2}\right)\right)^2 &= \int_0^\infty \xi^{-\frac{1}{2}} e^{-\xi} d\xi \int_0^\infty \zeta^{-\frac{1}{2}} e^{-\zeta} d\zeta \\ &= \int_0^\infty \int_0^\infty \xi^{-\frac{1}{2}} \zeta^{-\frac{1}{2}} e^{-(\xi+\zeta)} d\xi d\zeta \\ &= \int_0^{\pi/2} \int_0^\infty \frac{1}{r^2 \sin\theta \cos\theta} e^{-r^2} 4r^3 \sin\theta \cos\theta dr d\theta \\ &= \int_0^{\pi/2} \int_0^\infty e^{-r^2} 4r dr d\theta = \pi.\end{aligned}$$

Many common distributions are special cases of the gamma distribution, as seen in Figure 8.7.

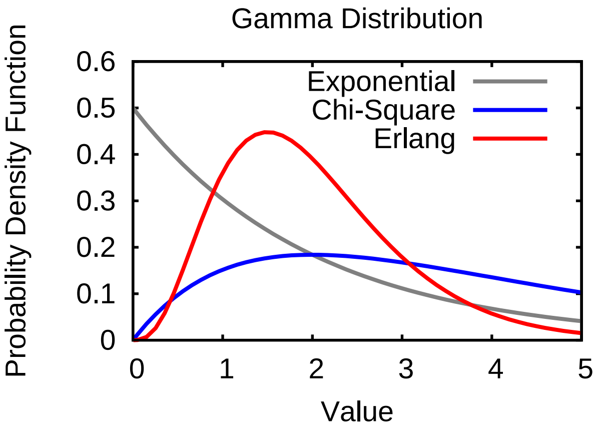

Figure 8.7: Gamma distributions form a two-parameter family of PDFs and, depending on (α, λ), they can be employed to model various situations. The parameters used above are $(1, 0.5)$ for the exponential distribution, $(2, 0.5)$ for the chi-square distribution and $(4, 2)$ for the Erlang distribution; they are all instances of gamma distributions.

The Exponential Distribution: When $\alpha = 1$, the gamma distribution simply reduces to the exponential distribution discussed in Section 8.3.3.

The Chi-Square Distribution: When $\lambda = 1/2$ and $\alpha = k/2$ for some positive integer k, the gamma distribution becomes a *chi-square distribution*,

$$f_X(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(k/2)} \quad x > 0.$$

The chi-square distribution is one of the probability distributions most widely used in statistical inference problems. Interestingly, the sum of the squares of k independent standard normal random variables leads to a chi-square variable with k degrees of freedom.

The Erlang Distribution: When $\alpha = m$, a positive integer, the gamma distribution is called an *Erlang distribution.* This distribution finds application in queueing theory. Its PDF is given by

$$f_X(x) = \frac{\lambda (\lambda x)^{m-1} e^{-\lambda x}}{(m-1)!} \quad x > 0.$$

An m-Erlang random variable can be obtained by summing m independent exponential random variables. Specifically, let $X_1, X_2, \ldots, X_m$ be independent exponential random variables, each with parameter $\lambda > 0$. The random variable S_m given by

$$S_m = \sum_{k=1}^{m} X_k$$

is an Erlang random variable with parameter m and λ.

Example 76. *Suppose that the requests arriving at a computer server on the Internet are characterized by independent, memoryless inter-arrival periods. Let S_m be a random variable that denotes the time instant of the mth arrival, then S_m is an Erlang random variable.*

8.4.2 The Rayleigh Distribution

The Rayleigh PDF is given by

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \quad r \geq 0.$$

The *Rayleigh distribution* arises in the context of wireless communications. Suppose that X and Y are two independent normal random variables, then

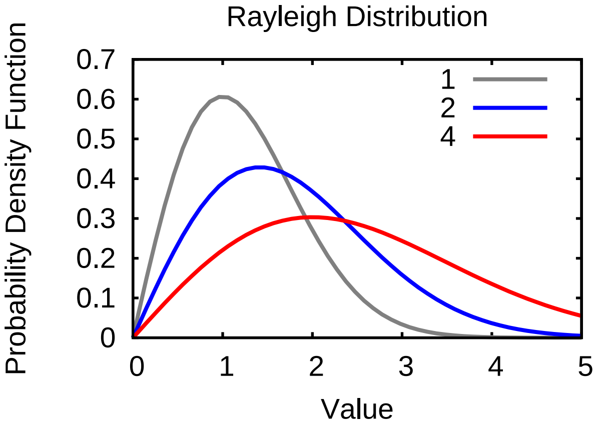

Figure 8.8: This figure plots the distributions of Rayleigh random variables for parameters $\sigma^2 \in \{1, 2, 4\}$.

the magnitude of this random vector possesses a Rayleigh distribution. Also, if R is a Rayleigh random variable then R^2 has an exponential distribution.

Example 77. *Radio signals propagating through wireless media get reflected, refracted and diffracted. This creates variations in signal strength at the destinations, a phenomenon known as fading. Rayleigh random variables are often employed to model amplitude fluctuations of radio signals in urban environments.*

8.4.3 The Laplace Distribution

The *Laplace distribution* is sometimes called a double exponential distribution because it can be thought of as an exponential function and its reflection spliced together. The PDF of a Laplacian random variable can then be written as

$$f_X(x) = \frac{1}{2b} e^{-\frac{|x|}{b}} \quad x \in \mathbb{R},$$

where b is a positive constant. The difference between two independent and identically distributed exponential random variables is governed by a Laplace distribution.

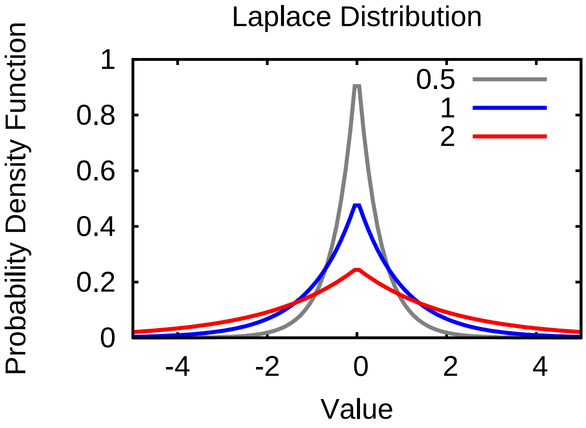

Figure 8.9: The PDF of a Laplace random variable can be constructed using an exponential function and its reflection spliced together. This figures shows Laplace PDFs for parameters $b \in \{0.5, 1, 2\}$.

8.4.4 The Cauchy Distribution

The *Cauchy distribution* is considered a heavy-tail distribution because its tail is not exponentially bounded. The PDF of a Cauchy random variable is given by

$$f_X(x) = \frac{\gamma}{\pi \left(\gamma^2 + x^2\right)} \quad x \in \mathbb{R}.$$

An interesting fact about this distribution is that its mean, variance and all higher moments are undefined. Moreover, if $X_1, X_2, \ldots, X_n$ are independent random variables, each with a standard Cauchy distribution, then the sample mean $(X_1 + X_2 + \cdots + X_n)/n$ possesses the same Cauchy distribution. Cauchy random variables appear in detection theory to model communication systems subject to extreme noise conditions; they also finds applications in physics. Physicists sometimes refer to this distribution as the *Lorentz distribution*.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Sections 5.1, 5.3–5.6.

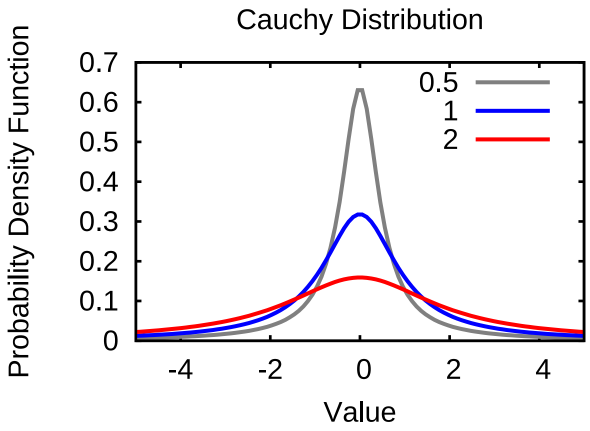

Figure 8.10: Cauchy distributions are categorized as heavy-tail distributions because of their very slow decay. The PDFs of Cauchy random variables are plotted above for parameters $\gamma \in \{0.5, 1, 2\}$.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Sections 3.1–3.3.

3. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 4.1, 5.1.

4. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Sections 3.1–3.4.

Chapter 9

Functions and Derived Distributions

We already know from our previous discussion that it is possible to form new random variables by applying real-valued functions to existing discrete random variables. In a similar manner, it is possible to generate a new random variable Y by taking a well-behaved function $g(\cdot)$ of a continuous random variable X. The graphical interpretation of this notion is analog to the discrete case and appears in Figure 9.1.

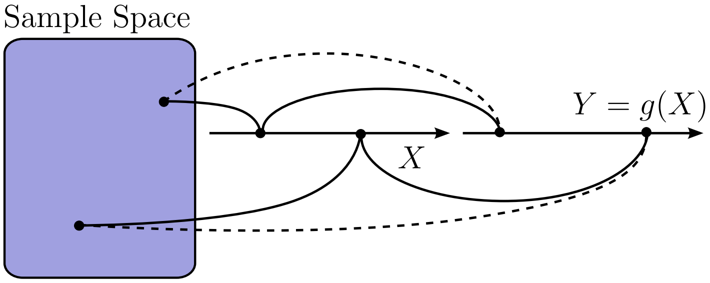

Figure 9.1: A function of a random variable is a random variable itself. In this figure, Y is obtained by applying function $g(\cdot)$ to the value of continuous random variable X.

Suppose X is a continuous random variable and let $g(\cdot)$ be a real-valued function. The function composition $Y = g(X)$ is itself a random variable. The probability that Y falls in a specific set S depends on both the function $g(\cdot)$

and the PDF of X,

$$\Pr(Y \in S) = \Pr(g(X) \in S) = \Pr\left(X \in g^{-1}(S)\right) = \int_{g^{-1}(S)} f_X(\xi) d\xi,$$

where $g^{-1}(S) = \{\xi \in X(\Omega) | g(\xi) \in S\}$ denotes the preimage of S. In particular, we can derive the CDF of Y using the formula

$$F_Y(y) = \Pr(g(X) \leq y) = \int_{\{\xi \in X(\Omega) | g(\xi) \leq y\}} f_X(\xi) d\xi. \tag{9.1}$$

Example 78. *Let X be a Rayleigh random variable with parameter $\sigma^2 = 1$, and define $Y = X^2$. We wish to find the distribution of Y. Using* (9.1)*, we can compute the CDF of Y. For $y > 0$, we get*

$$\begin{aligned}
F_Y(y) &= \Pr(Y \leq y) = \Pr\left(X^2 \leq y\right) \\
&= \Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_0^{\sqrt{y}} \xi e^{-\frac{\xi^2}{2}} d\xi \\
&= \int_0^y \frac{1}{2} e^{-\frac{\zeta}{2}} d\zeta = 1 - e^{-\frac{y}{2}}.
\end{aligned}$$

In this derivation, we use the fact that $X \geq 0$ in identifying the boundaries of integration, and we apply the change of variables $\zeta = \xi^2$ in computing the integral. We recognize $F_Y(\cdot)$ as the CDF of an exponential random variable. This shows that the square of a Rayleigh random variable possesses an exponential distribution.

In general, the fact that X is a continuous random variable does not provide much information about the properties of $Y = g(X)$. For instance, Y could be a continuous random variable, a discrete random variable or neither. To gain a better understanding of derived distributions, we begin our exposition of functions of continuous random variables by exploring specific cases.

9.1 Monotone Functions

A *monotonic function* is a function that preserves a given order. For instance, $g(\cdot)$ is monotone increasing if, for all x_1 and x_2 such that $x_1 \leq x_2$, we have $g(x_1) \leq g(x_2)$. Likewise, a function $g(\cdot)$ is termed monotone decreasing provided that $g(x_1) \geq g(x_2)$ whenever $x_1 \leq x_2$. If the inequalities above are replaced by strict inequalities ($<$ and $>$), then the corresponding functions are said to be *strictly monotonic*. Monotonic functions of random variables are straightforward to handle because they admit the simple characterization of their derived CDFs. For non-decreasing function $g(\cdot)$ of continuous random variable X, we have

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(g(X) \leq y) = \Pr(g(X) \in (-\infty, y]) \\ &= \Pr(X \in g^{-1}((-\infty, y])) = \Pr\left(X \leq \sup\left\{g^{-1}((-\infty, y])\right\}\right) \\ &= F_X\left(\sup\left\{g^{-1}((-\infty, y])\right\}\right). \end{aligned} \tag{9.2}$$

The supremum comes from the fact that multiple values of x may lead to a same value of y; that is, the preimage $g^{-1}(y) = \{x | g(x) = y\}$ may contain several elements. Furthermore, $g(\cdot)$ may be discontinuous and $g^{-1}(y)$ may not contain any value. These scenarios all need to be accounted for in our expression, and this is accomplished by selecting the largest value in the set $g^{-1}((-\infty, y])$.

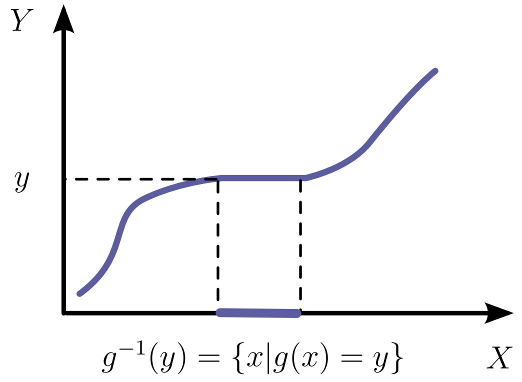

Figure 9.2: In this figure, Y is obtained by passing random variable X through a function $g(\cdot)$. The preimage of point y contains several elements, as seen above.

Example 79. *Let X be a continuous random variable uniformly distributed over interval $[0, 1]$. We wish to characterize the derived distribution of $Y = 2X$. This can be accomplished as follows. For $y \in [0, 2]$, we get*

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr\left(X \leq \frac{y}{2}\right) \\ &= \int_0^{\frac{y}{2}} dx = \frac{y}{2}. \end{aligned}$$

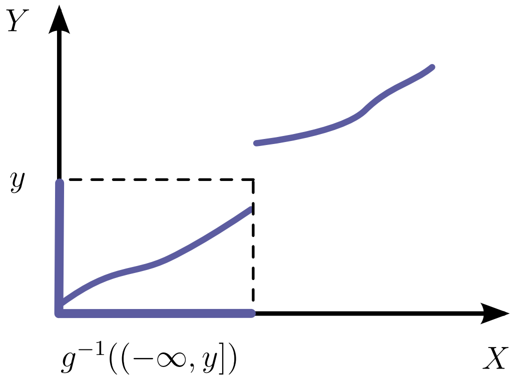

Figure 9.3: If $g(\cdot)$ is monotone increasing and discontinuous, then $g^{-1}(y)$ can be empty; whereas $g^{-1}((-\infty, y])$ is typically a well-defined interval. It is therefore advisable to define $F_Y(y)$ in terms of $g^{-1}((-\infty, y])$.

In particular, Y is a uniform random variable with support $[0, 2]$. By taking derivatives, we obtain the PDF of Y as

$$f_Y(y) = \begin{cases} \frac{1}{2}, & y \in [0, 2] \\ 0, & \textit{otherwise.} \end{cases}$$

More generally, an affine function of a uniform random variable is also a uniform random variable.

The same methodology applies to non-increasing functions. Suppose that $g(\cdot)$ is monotone decreasing, and let $Y = g(X)$ be a function of continuous random variable X. The CDF of Y is then equal to

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr\left(X \in g^{-1}((-\infty, y])\right) \\ &= \Pr\left(X \geq \inf\left\{g^{-1}((-\infty, y])\right\}\right) \\ &= 1 - F_X\left(\inf\left\{g^{-1}((-\infty, y])\right\}\right). \end{aligned} \tag{9.3}$$

This formula is similar to the previous case in that the infimum accounts for the fact that the preimage $g^{-1}(y) = \{x | g(x) = y\}$ may contain numerous elements or no elements at all.

9.2 Differentiable Functions

To further our understanding of derived distributions, we next consider the situation where $g(\cdot)$ is a differentiable and strictly increasing function. Note that, with these two properties, $g(\cdot)$ becomes an invertible function. It is therefore possible to write $x = g^{-1}(y)$ unambiguous, as the value of x is unique. In such a case, the CDF of $Y = g(X)$ becomes

$$F_Y(y) = \Pr\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right).$$

Differentiating this equation with respect to y, we obtain the PDF of Y

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(g^{-1}(y)\right) \\ &= f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) = f_X\left(g^{-1}(y)\right) \frac{dx}{dy}. \end{aligned}$$

With the simple substitution $x = g^{-1}(y)$, we get

$$f_Y(y) = f_X(x) \frac{dx}{dy} = \frac{f_X(x)}{\frac{dg}{dx}(x)}.$$

Note that $\frac{dg}{dx}(x) = \left|\frac{dg}{dx}(x)\right|$ is strictly positive because $g(\cdot)$ is a strictly increasing function. From this analysis, we gather that $Y = g(X)$ is a continuous random variable. In addition, we can express the PDF of $Y = g(X)$ in terms of the PDF of X and the derivative of $g(\cdot)$, as seen above.

Likewise, suppose that $g(\cdot)$ is differentiable and strictly decreasing. We can write the CDF of random variable $Y = g(X)$ as follows,

$$F_Y(y) = \Pr(g(X) \leq y) = \Pr\left(X \geq g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y)\right).$$

Its PDF is given by

$$f_Y(y) = \frac{d}{dy}\left(1 - F_X\left(g^{-1}(y)\right)\right) = \frac{f_X(x)}{-\frac{dg}{dx}(x)},$$

where again $x = g^{-1}(y)$. We point out that $\frac{dg}{dx}(x) = -\left|\frac{dg}{dx}(x)\right|$ is strictly negative because $g(\cdot)$ is a strictly decreasing function. As before, we find that $Y = g(X)$ is a continuous random variable and the PDF of Y can be expressed in terms of $f_X(\cdot)$ and the derivative of $g(\cdot)$. Combining these two expressions,

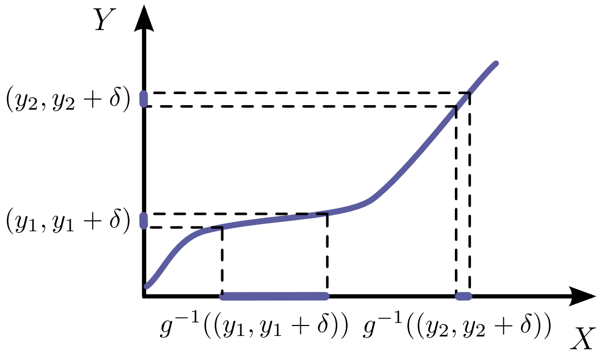

Figure 9.4: This figure provides a graphical interpretation of why the derivative of $g(\cdot)$ plays an important role in determining the value of the derived PDF $f_Y(\cdot)$. For an interval of width δ on the y-axis, the size of the corresponding interval on the x-axis depends heavily on the derivative of $g(\cdot)$. A small slope leads to a wide interval, whereas a steep slope produces a narrow interval on the x-axis.

we observe that, when $g(\cdot)$ is differentiable and strictly monotone, the PDF of Y becomes

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dx}{dy}\right| = \frac{f_X(x)}{\left|\frac{dg}{dx}(x)\right|} \tag{9.4}$$

where $x = g^{-1}(y)$. The role of $\left|\frac{dg}{dx}(\cdot)\right|$ in finding the derived PDF $f_Y(\cdot)$ is illustrated in Figure 9.4.

Example 80. *Suppose that X is a Gaussian random variable with PDF*

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

We wish to find the PDF of random variable Y where $Y = aX + b$ and $a \neq 0$.

In this example, we have $g(x) = ax + b$ and $g(\cdot)$ is immediately recognized as a strictly monotonic function. The inverse of function of $g(\cdot)$ is equal to

$$x = g^{-1}(y) = \frac{y-b}{a},$$

and the desired derivative is given by

$$\frac{dx}{dy} = \frac{1}{\frac{dg}{dx}(x)} = \frac{1}{a}.$$

The PDF of Y can be computed using (9.4), *and is found to be*

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dx}{dy}\right| = \frac{1}{\sqrt{2\pi}|a|}e^{-\frac{(y-b)^2}{2a^2}},$$

which is itself a Gaussian distribution.

Using a similar progression, we can show that the affine function of any Gaussian random variable necessarily remains a Gaussian random variable (provided $a \neq 0$).

Example 81 (Channel Fading and Energy)**.** *Suppose X is a Rayleigh random variable with parameter $\sigma^2 = 1$, and let $Y = X^2$. We wish to derive the distribution of random variable Y using the PDF of X.*

Recall that the distribution of Rayleigh random variable X is given by

$$f_X(x) = xe^{-\frac{x^2}{2}} \quad x \geq 0.$$

Since Y is the square of X, we have $g(x) = x^2$. Note that X is a non-negative random variable and $g(x) = x^2$ is strictly monotonic over $[0, \infty)$*. The PDF of Y is therefore found to be*

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{dg}{dx}(x)\right|} = \frac{f_X\left(\sqrt{y}\right)}{\left|\frac{dg}{dx}\left(\sqrt{y}\right)\right|} = \frac{\sqrt{y}}{2\sqrt{y}}e^{-\frac{y}{2}} = \frac{1}{2}e^{-\frac{y}{2}},$$

where $y \geq 0$. Thus, random variable Y possesses an exponential distribution with parameter $1/2$*. It may be instructive to compare this derivation with the steps outlined in Example 78.*

Finally, suppose that $g(\cdot)$ is a differentiable function with a finite number of local extrema. Then, $g(\cdot)$ is piecewise monotonic and we can write the PDF of $Y = g(X)$ as

$$f_Y(y) = \sum_{\{x \in X(\Omega) | g(x) = y\}} \frac{f_X(x)}{\left|\frac{dg}{dx}(x)\right|} \tag{9.5}$$

for (almost) all values of $y \in \mathbb{R}$. That is, $f_Y(y)$ is obtained by first identifying the values of x for which $g(x) = y$. The PDF of Y is then computed explicitly by finding the local contribution of each of these values to $f_Y(y)$ using the methodology developed above. This is accomplished by applying (9.4) repetitively to every value of x for which $g(x) = y$. It is certainly useful to compare (9.5) to its discrete equivalent (5.4), which is easier to understand and visualize.

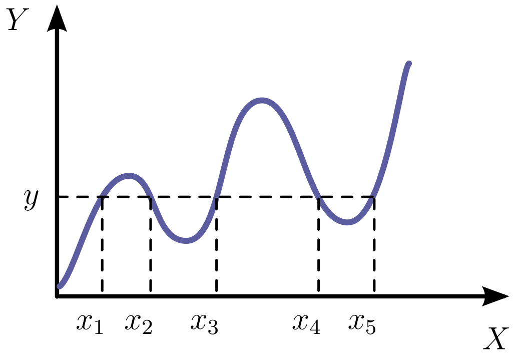

Figure 9.5: The PDF of $Y = g(X)$ when X is a continuous random variable and $g(\cdot)$ is differentiable with a finite number of local extrema is obtained by first identifying all the values of x for which $g(x) = y$, and then calculating the contribution of each of these values to $f_Y(y)$ using (9.4). The end result leads to (9.5).

Example 82 (Signal Phase and Amplitude)**.** *Suppose X is a continuous random variable uniformly distributed over $[0, 2\pi)$. Let $Y = \cos(X)$, the random sampling of a sinusoidal waveform. We wish to find the PDF of Y.*

For $y \in (-1, 1)$, the preimage $g^{-1}(y)$ contains two values in $[0, 2\pi)$, namely $\arccos(y)$ and $2\pi - \arccos(y)$. Recall that the derivative of $\cos(x)$ is given by

$$\frac{d}{dx} \cos(x) = -\sin(x).$$

Collecting these results, we can write the PDF of Y as

$$\begin{aligned} f_Y(y) &= \frac{f_X(\arccos(y))}{|-\sin(\arccos(y))|} + \frac{f_X(2\pi - \arccos(y))}{|-\sin(2\pi - \arccos(y))|} \\ &= \frac{1}{2\pi\sqrt{1-y^2}} + \frac{1}{2\pi\sqrt{1-y^2}} = \frac{1}{\pi\sqrt{1-y^2}}, \end{aligned}$$

where $-1 < y < 1$. The CDF of Y can be obtained by integrating $f_Y(y)$. Not surprisingly, solving this integral involves a trigonometric substitution.

9.3 Generating Random Variables

In many engineering projects, computer simulations are employed as a first step in validating concepts or comparing various design candidates. Many such tasks involve the generation of random variables. In this section, we explore a method to generate arbitrary random variables based on a routine that outputs a random value uniformly distributed between zero and one.

9.3.1 Continuous Random Variables

First, we consider a scenario where the simulation task requires the generation of a continuous random variable. We begin our exposition with a simple observation. Let X be a continuous random variable with PDF $f_X(\cdot)$. Consider the random variable $Y = F_X(X)$. Since $F_X(\cdot)$ is differentiable and strictly increasing over the support of X, we get

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{dF_X}{dx}(x)\right|} = \frac{f_X(x)}{|f_X(x)|} = 1$$

where $y \in (0,1)$ and $x = F_X^{-1}(y)$. The PDF of Y is zero outside of this interval because $0 \leq F_X(x) \leq 1$. Thus, using an arbitrary continuous random variable X, we can generate a uniform random variable Y with PDF

$$f_Y(y) = \begin{cases} 1 & y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

This observation provides valuable insight about our original goal. Suppose that Y is a continuous random variable uniformly distributed over $[0,1]$. We wish to generate continuous random variable with CDF $F_X(\cdot)$. First, we note that, when $F_X(\cdot)$ is invertible, we have

$$F_X^{-1}\left(F_X(X)\right) = X.$$

Thus, applying $F_X^{-1}(\cdot)$ to uniform random variable Y should lead to the desired result. Define $V = F_X^{-1}(Y)$, and consider the PDF of V. Using our knowledge of derived distributions, we get

$$f_V(v) = \frac{f_Y(y)}{\left|\frac{dF_X^{-1}}{dy}(y)\right|} = f_Y(y)\frac{dF_X}{dv}(v) = f_X(v)$$

where $y = F_X(v)$. Note that $f_Y(y) = 1$ for any $y \in [0,1]$ because Y is uniform over the unit interval. Hence the PDF of $F_X^{-1}(Y)$ possesses the structure wanted. We stress that this technique can be utilized to generate any random variable with PDF $f_X(\cdot)$ using a computer routine that outputs a random value uniformly distributed between zero and one. In other words, to create a continuous random variable X with CDF $F_X(\cdot)$, one can apply the function $F_X^{-1}(\cdot)$ to a random variable Y that is uniformly distributed over $[0, 1]$.

Example 83. *Suppose that Y is a continuous random variable uniformly distributed over $[0, 1]$. We wish to create an exponential random variable X with parameter λ by taking a function of Y.*

Random variable X is nonnegative, and its CDF is given by $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The inverse of $F_X(\cdot)$ is given by

$$F_X^{-1}(y) = -\frac{1}{\lambda} \log(1 - y).$$

We can therefore generate the desired random variable X with

$$X = -\frac{1}{\lambda} \log(1 - Y).$$

Indeed, for $x \geq 0$, we obtain

$$f_X(x) = \frac{f_Y(y)}{\frac{1}{\lambda(1-y)}} = \lambda e^{-\lambda x}$$

where we have implicitly defined $y = 1 - e^{-\lambda x}$. This is the desired distribution.

9.3.2 Discrete Random Variables

It is equally straightforward to generate a discrete random variable from a continuous random variable that is uniformly distributed between zero and one. Let $p_X(\cdot)$ be a PMF, and denote its support by $\{x_1, x_2, \ldots\}$ where $x_i < x_j$ whenever $i < j$. We know that the corresponding CDF is given by

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i).$$

We can generate a random variable X with PMF $p_X(\cdot)$ with the following case function,

$$g(y) = \begin{cases} x_i, & \text{if } F_X(x_{i-1}) < y \leq F_X(x_i) \\ 0, & \text{otherwise.} \end{cases}$$

Note that we have used the convention $x_0 = 0$ to simplify the definition of $g(\cdot)$. Taking $X = g(Y)$, we get

$$\begin{aligned}\Pr(X = x_i) &= \Pr(F_X(x_{i-1}) < Y \leq F_X(x_i)) \\ &= F_X(x_i) - F_X(x_{i-1}) = p_X(x_i).\end{aligned}$$

Of course, implementing a discrete random variable through a case statement may lead to an excessively slow routine. For many discrete random variables, there are much more efficient ways to generate a specific output.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Section 5.7.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 3.6.

3. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Section 4.6.

4. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 1.1,1.3–1.4.

5. Mitzenmacher, M., and Upfal, E., *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*, Cambridge, 2005: Chapters 1 & 10.

Chapter 10

Expectations and Bounds

The concept of expectation, which was originally introduced in the context of discrete random variables, can be generalized to other types of random variables. For instance, the expectation of a continuous random variable is defined in terms of its probability density function (PDF). We know from our previous discussion that expectations provide an effective way to summarize the information contained in the distribution of a random variable. As we will see shortly, expectations are also very valuable in establishing bounds on probabilities.

10.1 Expectations Revisited

The definition of an expectation associated with a continuous random variable is very similar to its discrete counterpart; the weighted sum is simply replaced by a weighted integral. For a continuous random variable X with PDF $f_X(\cdot)$, the *expectation* of $g(X)$ is defined by

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(\xi) f_X(\xi) d\xi.$$

In particular, the *mean* of X is equal to

$$\mathrm{E}[X] = \int_{\mathbb{R}} \xi f_X(\xi) d\xi$$

and its *variance* becomes

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \int_{\mathbb{R}} (\xi - \mathrm{E}[X])^2 f_X(\xi) d\xi.$$

As before, the variance of random variable X can also be computed using $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (E[X])^2$.

Example 84. *We wish to calculate the mean and variance of a Gaussian random variable with parameters m and σ^2. By definition, the PDF of this random variable can be written as*

$$f_X(\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi-m)^2}{2\sigma^2}} \quad \xi \in \mathbb{R}.$$

The mean of X can be obtained through direct integration, with a change of variables,

$$\begin{aligned} \mathrm{E}[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \xi e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\zeta + \frac{m}{\sigma}\right) e^{-\frac{\zeta^2}{2}} d\zeta \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \zeta e^{-\frac{\zeta^2}{2}} d\zeta + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{m}{\sigma} e^{-\frac{\zeta^2}{2}} d\zeta = m. \end{aligned}$$

In finding a solution, we have leveraged the facts that $\zeta e^{-\frac{\zeta^2}{2}}$ is an absolutely integrable, odd function. We also took advantage of the normalization condition which ensures that a Gaussian PDF integrates to one. To derive the variance, we again use the normalization condition. For a Gaussian PDF, this property implies that

$$\int_{-\infty}^{\infty} e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi = \sqrt{2\pi}\sigma.$$

Differentiating both sides of this equation with respect to σ, we get

$$\int_{-\infty}^{\infty} \frac{(\xi-m)^2}{\sigma^3} e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi = \sqrt{2\pi}.$$

Rearranging the terms yields

$$\int_{-\infty}^{\infty} \frac{(\xi-m)^2}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi = \sigma^2.$$

Hence, $\mathrm{Var}(X) = \mathrm{E}\left[(X-m)^2\right] = \sigma^2$. *Of course, the variance can also be obtained by more conventional methods.*

Example 85. *Suppose that R is a Rayleigh random variable with parameter σ^2. We wish to compute its mean and variance.*

Recall that R is a nonnegative random variable with PDF

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \quad r \geq 0.$$

Using this distribution, we get

$$\begin{aligned} \mathrm{E}[R] &= \int_0^\infty \xi f_R(\xi) d\xi = \int_0^\infty \frac{\xi^2}{\sigma^2} e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= -\xi e^{-\frac{\xi^2}{2\sigma^2}} \Big|_0^\infty + \int_0^\infty e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= \sqrt{2\pi}\sigma \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\zeta^2}{2\sigma^2}} d\zeta = \frac{\sqrt{2\pi}\sigma}{2}. \end{aligned}$$

Integration by parts is key in solving this expectation. Also, notice the judicious use of the fact that the integral of a standard normal random variable over $[0, \infty)$ must be equal to $1/2$. We compute the second moment of R below,

$$\begin{aligned} \mathrm{E}\left[R^2\right] &= \int_0^\infty \xi^2 f_R(\xi) d\xi = \int_0^\infty \frac{\xi^3}{\sigma^2} e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= -\xi^2 e^{-\frac{\xi^2}{2\sigma^2}} \Big|_0^\infty + \int_0^\infty 2\xi e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= -2\sigma^2 e^{-\frac{\xi^2}{2\sigma^2}} \Big|_0^\infty = 2\sigma^2. \end{aligned}$$

The variance of R is therefore equal to

$$\mathrm{Var}[R] = \frac{(4-\pi)}{2} \sigma^2.$$

Typically, σ^2 is employed to denote the variance of a random variable. It may be confusing at first to have a random variable R described in terms of parameter σ^2 whose variance is equal to $(4-\pi)\sigma^2/2$. This situation is an artifact of the following relation. A Rayleigh random variable R can be generated through the expression $R = \sqrt{X^2 + Y^2}$, where X and Y are independent zero-mean Gaussian variables with variance σ^2. Thus, the parameter σ^2 in $f_R(\cdot)$ is a tribute to this popular construction, not a representation of its actual variance.

For nonnegative random variable X, an alternative way to compute $\mathrm{E}[X]$ is described in Proposition 7.

Proposition 7. *Suppose that X is a nonnegative random variable with finite mean, then*

$$\mathrm{E}[X] = \int_0^\infty \Pr(X > x) dx.$$

Proof. We offer a proof for the special case where X is a continuous random variable, although the result remains true in general,

$$\begin{aligned}\int_0^\infty \Pr(X > x) dx &= \int_0^\infty \int_x^\infty f_X(\xi) d\xi dx \\ &= \int_0^\infty \int_0^\xi f_X(\xi) dx d\xi \\ &= \int_0^\infty \xi f_X(\xi) d\xi = \mathrm{E}[X].\end{aligned}$$

Interchanging the order of integration is justified because X is assumed to have finite mean. □

Example 86. *A player throws darts at a circular target hung on a wall. The dartboard has unit radius, and the position of every dart is distributed uniformly over the target. We wish to compute the expected distance from each dart to the center of the dartboard.*

Let R denote the distance from a dart to the center of the target. For $0 \leq r \leq 1$, the probability that R exceeds r is given by

$$\Pr(R > r) = 1 - \Pr(R \leq r) = 1 - \frac{\pi r^2}{\pi} = 1 - r^2.$$

Then, by Proposition 7, the expected value of R is equal to

$$\mathrm{E}[R] = \int_0^1 \left(1 - r^2\right) dr = \left(r - \frac{r^3}{3} \right) \Bigg|_0^1 = 1 - \frac{1}{3} = \frac{2}{3}.$$

Notice how we were able to compute the answer without deriving an explicit expression for $f_R(\cdot)$.

10.2 Moment Generating Functions

The *moment generating function* of a random variable X is defined by

$$M_X(s) = \mathrm{E}\left[e^{sX}\right].$$

For continuous random variables, the moment generating function becomes

$$M_X(s) = \int_{-\infty}^{\infty} f_X(\xi) e^{s\xi} d\xi .$$

The experienced reader will quickly recognize the definition of $M_X(s)$ as a variant of the *Laplace Transform*, a widely used linear operator. The moment generating function gets its name from the following property. Suppose that $M_X(s)$ exists within an open interval around $s = 0$, then the nth moment of X is given by

$$\begin{aligned} \frac{d^n}{ds^n} M_X(s) \Big|_{s=0} &= \frac{d^n}{ds^n} \mathrm{E}\left[e^{sX} \right] \Big|_{s=0} = \mathrm{E}\left[\frac{d^n}{ds^n} e^{sX} \right] \Big|_{s=0} \\ &= \mathrm{E}\left[X^n e^{sX} \right] \Big|_{s=0} = \mathrm{E}[X^n]. \end{aligned}$$

In words, if we differentiate $M_X(s)$ a total of n times and then evaluate the resulting function at zero, we obtain the nth moment of X. In particular, we have $\frac{dM_X}{ds}(0) = \mathrm{E}[X]$ and $\frac{d^2 M_X}{ds^2}(0) = \mathrm{E}[X^2]$.

Example 87 (Exponential Random Variable)**.** *Let X be an exponential random variable with parameter λ. The moment-generating function of X is given by*

$$M_X(s) = \int_0^{\infty} \lambda e^{-\lambda\xi} e^{s\xi} d\xi = \int_0^{\infty} \lambda e^{-(\lambda - s)\xi} d\xi = \frac{\lambda}{\lambda - s}.$$

The mean of X is

$$\mathrm{E}[X] = \frac{dM_X}{ds}(0) = \frac{\lambda}{(\lambda - s)^2} \Big|_{s=0} = \frac{1}{\lambda};$$

more generally, the nth moment of X can be computed as

$$\mathrm{E}[X^n] = \frac{d^n M_X}{ds^n}(0) = \frac{n! \lambda}{(\lambda - s)^{n+1}} \Big|_{s=0} = \frac{n!}{\lambda^n}.$$

Incidentally, we can deduce from these results that the variance of X is $1/\lambda^2$.

The definition of the moment generating function applies to discrete random variables as well. In fact, for integer-valued random variables, the moment generating function and the ordinary generating function are related through the equation

$$M_X(s) = \sum_{k \in X(\Omega)} e^{sk} p_X(k) = G_X(e^s).$$

Example 88 (Discrete Uniform Random Variable)**.** *Suppose U is a discrete uniform random variable taking value in $U(\Omega) = \{1, 2, \ldots, n\}$. Then, $p_U(k) = 1/n$ for $1 \leq k \leq n$ and*

$$M_U(s) = \sum_{k=1}^{n} \frac{1}{n} e^{sk} = \frac{1}{n} \sum_{k=1}^{n} e^{sk} = \frac{e^s(e^{ns} - 1)}{n(e^s - 1)}.$$

The moment generating function provides an alternate and somewhat intricate way to compute the mean of U,

$$\begin{aligned} \mathrm{E}[U] = \frac{dM_U}{ds}(0) &= \lim_{s \to 0} \frac{ne^{(n+2)s} - (n+1)e^{(n+1)s} + e^s}{n\left(e^s - 1\right)^2} \\ &= \lim_{s \to 0} \frac{n(n+2)e^{(n+1)s} - (n+1)^2 e^{ns} + 1}{2n\left(e^s - 1\right)} \\ &= \lim_{s \to 0} \frac{n(n+1)(n+2)ne^{(n+1)s} - n(n+1)^2 e^{ns}}{2ne^s} = \frac{n+1}{2}. \end{aligned}$$

Notice the double application of l'Hôpital's rule *to evaluate the derivative of $M_U(s)$ at zero. This may be deemed a more contrived method to derive the expected value of a discrete uniform random variables, but it does not rely on prior knowledge of special sums. Through similar steps, one can derive the second moment of U, which is equal to*

$$\mathrm{E}\left[U^2\right] = \frac{(n+1)(2n+1)}{6}.$$

From these two results, we can show that the variance of U is $(n^2 - 1)/12$.

The simple form of the moment generating function of a standard normal random variable points to its importance in many situations. The exponential function is analytic and possesses many representations.

Example 89 (Gaussian Random Variable)**.** *Let X be a standard normal random variable whose PDF is given by*

$$f_X(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}.$$

The moment generating function of X is equal to

$$\begin{aligned} M_X(s) = \mathrm{E}\left[e^{sX}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} e^{s\xi} d\xi \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2 + 2s\xi}{2}} d\xi = e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2 - 2s\xi + s^2}{2}} d\xi \\ &= e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\xi - s)^2}{2}} d\xi = e^{\frac{s^2}{2}}. \end{aligned}$$

The last equality follows from the normalization condition and the fact that the integrand is a Gaussian PDF.

Let $M_X(s)$ be the moment generating function associated with a random variable X, and consider the random variable $Y = aX + b$ where a and b are constant. The moment generating function of Y can be obtained as follows,

$$M_Y(s) = \mathrm{E}\left[e^{sX}\right] = \mathrm{E}\left[e^{s(aX+b)}\right] = e^{sb}\mathrm{E}\left[e^{saX}\right] = e^{sb}M_X(as).$$

Thus, if Y is an affine function of X then $M_Y(s) = e^{sb}M_X(as)$.

Example 90. *We can use this property to identify the moment generating function of a Gaussian random variable with parameters m and σ^2. Recall that an affine function of a Gaussian random variable is also Gaussian. Let $Y = \sigma X + m$, then the moment generating function of Y becomes*

$$M_Y(s) = \mathrm{E}\left[e^{sY}\right] = \mathrm{E}\left[e^{s(\sigma X+m)}\right] = e^{sm}\mathrm{E}\left[e^{s\sigma X}\right] = e^{sm+\frac{s^2\sigma^2}{2}}.$$

From this moment generating function, we get

$$\begin{aligned}
\mathrm{E}[Y] &= \frac{dM_Y}{ds}(0) = \left[\left(m + s\sigma^2\right) e^{sm+\frac{s^2\sigma^2}{2}}\right]\Big|_{s=0} = m \\
\mathrm{E}\left[Y^2\right] &= \frac{d^2M_Y}{ds^2}(0) = \left[\sigma^2 e^{sm+\frac{s^2\sigma^2}{2}} + \left(m + s\sigma^2\right)^2 e^{sm+\frac{s^2\sigma^2}{2}}\right]\Big|_{s=0} \\
&= \sigma^2 + m^2.
\end{aligned}$$

The mean of Y is m and its variance is equal to σ^2, as anticipated.

10.3 Important Inequalities

There are many situations for which computing the exact value of a probability is impossible or impractical. In such cases, it may be acceptable to provide bounds on the value of an elusive probability. The expectation is most important in finding pertinent bounds.

As we will see, many upper bounds rely on the concept of dominating functions. Suppose that $g(x)$ and $h(x)$ are two nonnegative function such that $g(x) \leq h(x)$ for all $x \in \mathbb{R}$. Then, for any continuous random variable X, the following inequality holds

$$\begin{aligned}\mathrm{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &\leq \int_{-\infty}^{\infty} h(x) f_X(x) dx = \mathrm{E}[h(X)].\end{aligned}$$

This is illustrated in Figure 10.1. In words, the weighted integral of $g(\cdot)$ is dominated by the weighted integral of $h(\cdot)$, where $f_X(\cdot)$ acts as the weighting function. This notion is instrumental in understanding bounding techniques.

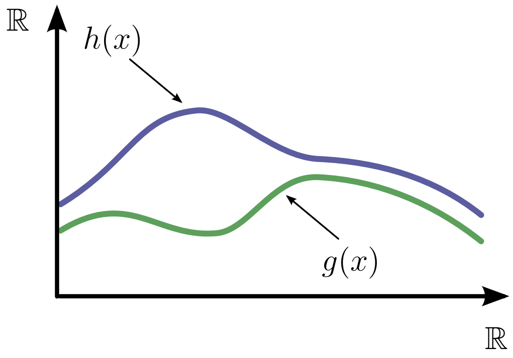

Figure 10.1: If $g(x)$ and $h(x)$ are two nonnegative functions such that $g(x) \leq h(x)$ for all $x \in \mathbb{R}$, then $\mathrm{E}[g(X)]$ is less than or equal to $\mathrm{E}[h(X)]$.

10.3.1 The Markov Inequality

We begin our exposition of classical upper bounds with a result known as the *Markov inequality*. Recall that, for admissible set $S \subset \mathbb{R}$, we have

$$\Pr(X \in S) = \mathrm{E}\left[\mathbf{1}_S(X)\right].$$

Thus, to obtain a bound on $\Pr(X \in S)$, it suffices to find a function that dominates $\mathbf{1}_S(\cdot)$ and for which we can compute the expectation.

Suppose that we wish to bound $\Pr(X \geq a)$ where X is a nonnegative random variable. In this case, we can select $S = [a, \infty)$ and function $h(x) = x/a$. For any $x \geq 0$, we have $h(x) \geq \mathbf{1}_S(x)$, as illustrated in Figure 10.2. It follows that

$$\Pr(X \geq a) = \mathrm{E}\left[\mathbf{1}_S(X)\right] \leq \frac{\mathrm{E}[X]}{a}.$$

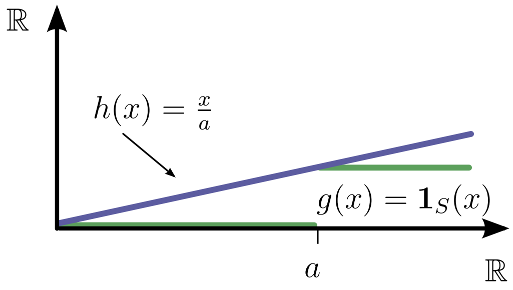

Figure 10.2: Suppose that we wish to find a bound for $\Pr(X \leq a)$. We define set $S = [a, \infty)$ and function $g(x) = \mathbf{1}_S(x)$. Using dominating function $h(x) = x/a$, we conclude that $\Pr(X \geq a) \leq a^{-1}\mathrm{E}[X]$ for any nonnegative random variable X.

10.3.2 The Chebyshev Inequality

The *Chebyshev inequality* provides an extension to this methodology to various dominating functions. This yields a number of bounds that become useful in a myriad of contexts.

Proposition 8 (Chebyshev Inequality). *Suppose $h(\cdot)$ is a nonnegative function and let S be an admissible set. We denote the infimum of $h(\cdot)$ over S by*

$$i_S = \inf_{x \in S} h(x).$$

The Chebyshev inequality asserts that

$$i_S \Pr(X \in S) \leq \mathrm{E}[h(X)] \tag{10.1}$$

where X is an arbitrary random variable.

Proof. This is a remarkably powerful result and it can be shown in a few steps. The definition of i_S and the fact that $h(\cdot)$ is nonnegative imply that

$$i_S \mathbf{1}_S(x) \leq h(x) \mathbf{1}_S(x) \leq h(x)$$

for any $x \in \mathbb{R}$. Moreover, for any such x and distribution $f_X(\cdot)$, we can write $i_S \mathbf{1}_S(x) f_X(x) \leq h(x) f(x)$, which in turn yields

$$\begin{aligned} i_S \Pr(X \in S) &= \mathrm{E}\left[i_s \mathbf{1}_S(X) \right] = \int_{\mathbb{R}} i_S \mathbf{1}_S(\xi) f_X(\xi) d\xi \\ &\leq \int_{\mathbb{R}} h(\xi) f_X(\xi) d\xi = \mathrm{E}[h(X)]. \end{aligned}$$

When $i_S > 0$, this provides the upper bound $\Pr(X \in S) \leq i_S^{-1} \mathrm{E}[h(X)]$. □

Although the proof assumes a continuous random variable, we emphasize that the Chebyshev inequality applies to both discrete and continuous random variables alike. The interested reader can rework the proof using the discrete setting and a generic PMF. We provide special instances of the Chebyshev inequality below.

Example 91. *Consider the nonnegative function $h(x) = x^2$ and let $S = \{x | x^2 \geq b^2\}$ where b is a positive constant. We wish to find a bound on the probability that $|X|$ exceeds b. Using the Chebyshev inequality, we have $i_S = \inf_{x \in S} x^2 = b^2$ and, consequently, we get*

$$b^2 \Pr(X \in S) \leq \mathrm{E}\left[X^2\right].$$

Constant b being a positive real number, we can rewrite this equation as

$$\Pr(|X| \geq b) = \Pr(X \in S) \leq \frac{\mathrm{E}\left[X^2\right]}{b^2}.$$

Example 92 (The Cantelli Inequality). *Suppose that X is a random variable with mean m and variance σ^2. We wish to show that*

$$\Pr(X - m \geq a) \leq \frac{\sigma^2}{a^2 + \sigma^2},$$

where $a \geq 0$.

This equation is slightly more involved and requires a small optimization in addition to the Chebyshev inequality. Define $Y = X - m$ and note that, by construction, we have $\mathrm{E}[Y] = 0$. Consider the probability $\Pr(Y \geq a)$ where $a > 0$, and let $S = \{y | y \geq a\}$. Also, define the nonnegative function $h(y) = (y+b)^2$, where $b > 0$. Following the steps of the Chebyshev inequality, we write the infimum of $h(y)$ over S as

$$i_S = \inf_{y \in S} (y + b)^2 = (a + b)^2.$$

Then, applying the Chebyshev inequality, we obtain

$$\Pr(Y \geq a) \leq \frac{\mathrm{E}\left[(Y + b)^2\right]}{(a + b)^2} = \frac{\sigma^2 + b^2}{(a + b)^2}. \quad (10.2)$$

This inequality holds for any $b > 0$. To produce a better upper bound, we minimize the right-hand side of (10.2) *over all possible values of b. Differentiating this expression and setting the derivative equal to zero yields*

$$\frac{2b}{(a+b)^2} = \frac{2\left(\sigma^2 + b^2\right)}{(a+b)^3}$$

or, equivalently, $b = \sigma^2/a$. A second derivative test reveals that this is indeed a minimum. Collecting these results, we obtain

$$\Pr(Y \geq a) \leq \frac{\sigma^2 + b^2}{(a+b)^2} = \frac{\sigma^2}{a^2 + \sigma^2}.$$

Substituting $Y = X - m$ leads to the desired result.

In some circumstances, a Chebyshev inequality can be tight.

Example 93. *Let a and b be two constants such that $0 < b \leq a$. Consider the function $h(x) = x^2$ along with the set $S = \{x | x^2 \geq a^2\}$. Furthermore, let X be a discrete random variable with PMF*

$$p_X(x) = \begin{cases} 1 - \frac{b^2}{a^2}, & x = 0 \\ \frac{b^2}{a^2}, & x = a \\ 0, & \textit{otherwise.} \end{cases}$$

For this random variable, we have $\Pr(X \in S) = b^2/a^2$. By inspection, we also gather that the second moment of X is equal to $\mathrm{E}\left[X^2\right] = b^2$. Applying the Chebyshev inequality, we get $i_S = \inf_{x \in S} h(x) = a^2$ and therefore

$$\Pr(X \in S) \leq i_S^{-1} \mathrm{E}\left[h(X)\right] = \frac{b^2}{a^2}.$$

Thus, in this particular example, the inequality is met with equality.

10.3.3 The Chernoff Bound

The *Chernoff bound* is yet another upper bound that can be constructed from the Chebyshev inequality. Still, because of its central role in many application domains, it deserves its own section. Suppose that we want to find a bound on the probability $\Pr(X \geq a)$. We can apply the Chebyshev inequality using the nonnegative function $h(x) = e^{sx}$, where $s > 0$. For this specific construction, $S = [a, \infty)$ and

$$i_S = \inf_{x \in S} e^{sx} = e^{sa}.$$

It follows that

$$\Pr(X \geq a) \leq e^{-sa} \mathrm{E}[e^{sX}] = e^{-sa} M_X(s).$$

Because this inequality holds for any $s > 0$, we can optimize the upper bound over all possible values of s, thereby picking the best one,

$$\Pr(X \geq a) \leq \inf_{s > 0} e^{-sa} M_X(s). \tag{10.3}$$

This inequality is called the Chernoff bound. It is sometimes expressed in terms of the log-moment generating function $\Lambda(s) = \log M_X(s)$. In this latter case, (10.3) translates into

$$\log \Pr(X \geq a) \leq - \sup_{s > 0} \left\{ sa - \Lambda(s) \right\}. \tag{10.4}$$

The right-hand side of (10.4) is called the *Legendre transformation* of $\Lambda(s)$. Figure 10.3 plots $e^{s(x-a)}$ for various values of $s > 0$. It should be noted that all these functions dominate $\mathbf{1}_{[a,\infty)}(x)$, and therefore they each provide a different bound on $\Pr(X \geq a)$. It is natural to select the function that provides the best bound. Yet, in general, this optimal $e^{s(x-a)}$ may depend on the distribution of X and the value of a, which explains why (10.3) involves a search over all possible values of s.

10.3.4 Jensen's Inequality

Some inequalities can be derived based on the properties of a single function. The *Jensen inequality* is one such example. Suppose that function $g(\cdot)$ is convex and twice differentiable, with

$$\frac{d^2 g}{dx^2}(x) \geq 0$$

for all $x \in \mathbb{R}$. From the fundamental theorem of calculus, we have

$$g(x) = g(a) + \int_a^x \frac{dg}{dx}(\xi) d\xi.$$

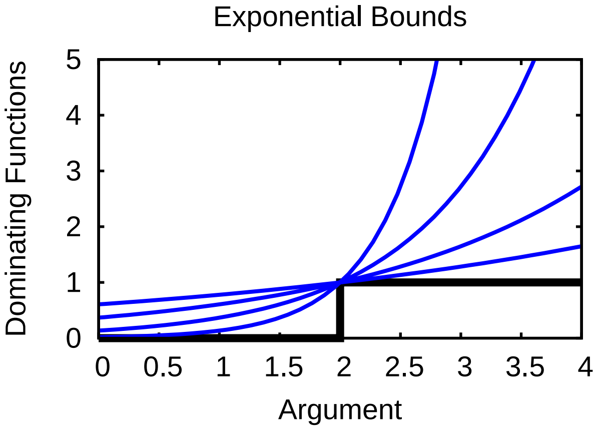

Figure 10.3: This figure illustrates how exponential functions can be employed to provide bounds on $\Pr(X > a)$. Optimizing over all admissible exponential functions, $e^{s(x-a)}$ where $s > 0$, leads to the celebrated Chernoff bound.

Futhermore, because the second derivative of $g(\cdot)$ is a non-negative function, we gather that $\frac{dg}{dx}(\cdot)$ is a monotone increasing function. As such, for any value of a, we have

$$\begin{aligned} g(x) &= g(a) + \int_a^x \frac{dg}{dx}(\xi) d\xi \\ &\geq g(a) + \int_a^x \frac{dg}{dx}(a) d\xi = g(a) + (x-a)\frac{dg}{dx}(a). \end{aligned}$$

For random variable X, we then have

$$g(X) \geq g(a) + (X-a)\frac{dg}{dx}(a).$$

Choosing $a = \mathrm{E}[X]$ and taking expectations on both sides, we obtain

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]) + (\mathrm{E}[X] - \mathrm{E}[X])\frac{dg}{dx}(\mathrm{E}[X]) = g(\mathrm{E}[X]).$$

That is, $\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$, provided that these two expectations exist. The Jensen inequality actually holds for convex functions that are not twice differentiable, but the proof is much harder in the general setting.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Sections 5.2, 7.7, 8.2.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 3.1, 4.1, 7.1.

3. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 4.2–4.3.

4. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Section 5.2.

Chapter 11

Multiple Continuous Random Variables

Being versed at dealing with multiple random variables is an essential part of statistics, engineering and science. This is equally true for models based on discrete and continuous random variables. In this chapter, we focus on the latter and expand our exposition of continuous random variables to random vectors. Again, our initial survey of this topic revolves around conditional distributions and pairs of random variables. More complex scenarios will be considered in the later parts of the chapter.

11.1 Joint Cumulative Distributions

Let X and Y be two random variables associated with a same experiment. The *joint cumulative distribution function* of X and Y is defined by

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) \quad x, y \in \mathbb{R}.$$

Keeping in mind that X and Y are real-valued functions acting on a same sample space, we can also write

$$F_{X,Y}(x,y) = \Pr\left(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}\right).$$

From this characterization, we can identify a few properties of the joint CDF;

$$\begin{aligned}\lim_{y \uparrow \infty} F_{X,Y}(x,y) &= \lim_{y \uparrow \infty} \Pr\left(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}\right) \\ &= \Pr\left(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \in \mathbb{R}\}\right) \\ &= \Pr\left(\{\omega \in \Omega | X(\omega) \leq x\}\right) = F_X(x).\end{aligned}$$

Similarly, we have $\lim_{x \uparrow \infty} F_{X,Y}(x,y) = F_Y(y)$. Taking limits in the other direction, we get

$$\lim_{x \downarrow -\infty} F_{X,Y}(x,y) = \lim_{y \downarrow -\infty} F_{X,Y}(x,y) = 0.$$

When the function $F_{X,Y}(\cdot,\cdot)$ is totally differentiable, it is possible to define the *joint probability density function* of X and Y,

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial y \partial x}(x,y) \quad x, y \in \mathbb{R}. \tag{11.1}$$

Hereafter, we refer to a pair of random variables as continuous if the corresponding joint PDF exists and is defined unambiguously through (11.1). When this is the case, standard calculus asserts that the following equation holds,

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(\xi,\zeta) d\zeta d\xi.$$

From its definition, we note that $f_{X,Y}(\cdot,\cdot)$ is a nonnegative function which integrates to one,

$$\iint_{\mathbb{R}^2} f_{X,Y}(\xi,\zeta) d\zeta d\xi = 1.$$

Furthermore, for any admissible set $S \subset \mathbb{R}^2$, the probability that $(X,Y) \in S$ can be evaluated through the integral formula

$$\begin{aligned}\Pr((X,Y) \in S) &= \iint_{\mathbb{R}^2} \mathbf{1}_S(\xi,\zeta) f_{X,Y}(\xi,\zeta) d\zeta d\xi \\ &= \iint_{S} f_{X,Y}(\xi,\zeta) d\zeta d\xi.\end{aligned} \tag{11.2}$$

In particular, if S is the cartesian product of two intervals,

$$S = \left\{ (x,y) \in \mathbb{R}^2 \middle| a \leq x \leq b, c \leq y \leq d \right\},$$

then the probability that $(X, Y) \in S$ reduces to the typical integral form

$$\Pr((X, Y) \in S) = \Pr(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(\xi, \zeta) d\zeta d\xi.$$

Example 94. *Suppose that the random pair (X, Y) is uniformly distributed over the unit circle. We can express the joint PDF $f_{X,Y}(\cdot, \cdot)$ as*

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & \textit{otherwise.} \end{cases}$$

We wish to find the probability that the point (X, Y) lies inside a circle of radius $1/2$.

Let $S = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 \leq 0.5\}$. The probability that (X, Y) belongs to S is given by

$$\Pr((X, Y) \in S) = \iint_{\mathbb{R}^2} \frac{\mathbf{1}_S(\xi, \zeta)}{\pi} d\xi d\zeta = \frac{1}{4}.$$

Thus, the probability that (X, Y) is contained within a circle of radius half is one fourth.

Example 95. *Let X and Y be two independent zero-mean Gaussian random variables, each with variance σ^2. For $(x, y) \in \mathbb{R}^2$, their joint PDF is given by*

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

We wish to find the probability that (X, Y) falls within a circle of radius r centered at the origin.

We can compute this probability using integral formula (11.2) applied to this particular problem. Let $R = \sqrt{X^2 + Y^2}$ and assume $r > 0$, then

$$\begin{aligned} \Pr(R \leq r) &= \iint_{R \leq r} f_{X,Y}(x, y) dx dy = \iint_{R \leq r} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}(x, y) dx dy \\ &= \int_0^r \int_0^{2\pi} \frac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}} r d\theta dr = 1 - e^{-\frac{r^2}{2\sigma^2}}. \end{aligned}$$

The probability that (X, Y) is contained within a circle of radius r is $1 - e^{-\frac{r^2}{2\sigma^2}}$. Recognizing that R is a continuous random variable, we can write its PDF as

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \quad r \geq 0.$$

From this equation, we gather That R possesses a Rayleigh distribution with parameter σ^2.

11.2 Conditional Probability Distributions

Given non-vanishing event A, we can write the conditional CDF of random variable X as

$$F_{X|A}(x) = \Pr(X \leq x|A) = \frac{\Pr(\{X \leq x\} \cap A)}{\Pr(A)} \quad x \in \mathbb{R}.$$

Note that event A can be defined in terms of variables X and Y. For instance, we may use $A = \{Y \geq X\}$ as our condition. Under suitable conditions, it is equally straightforward to specify the conditional PDF of X given A,

$$f_{X|A}(x) = \frac{dF_{X|A}}{dx}(x) \quad x \in \mathbb{R}.$$

Example 96. *Let X and Y be continuous random variables with joint PDF*

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} \quad x, y \geq 0.$$

We wish to compute the conditional PDF of X given $A = \{X \leq Y\}$. To solve this problem, we first compute the probability of the event $\{X \leq x\} \cap A$,

$$\begin{aligned}\Pr(\{X \leq x\} \cap A) &= \int_0^x \int_0^\xi f_{X,Y}(\xi,\zeta) d\zeta d\xi = \int_0^x \int_0^\xi \lambda^2 e^{-\lambda(\xi+\zeta)} d\zeta d\xi \\ &= \int_0^x \lambda e^{-\lambda\xi}\left(1 - e^{-\lambda\xi}\right) d\xi = \frac{\left(1 - e^{-\lambda x}\right)^2}{2}.\end{aligned}$$

By symmetry, we gather that $\Pr(A) = 1/2$ *and, as such,*

$$f_{X|A}(x) = 2\lambda e^{-\lambda x}\left(1 - e^{-\lambda x}\right) \quad x \geq 0.$$

One case of special interest is the situation where event A is defined in terms of the random variable X itself. In particular, consider the PDF of X conditioned on the fact that X belongs to an interval I. Then, $A = \{X \in I\}$ and the conditional CDF of X becomes

$$\begin{aligned}F_{X|A}(x) &= \Pr(X \leq x|X \in I) \\ &= \frac{\Pr\left(\{X \leq x\} \cap \{X \in I\}\right)}{\Pr(X \in I)} \\ &= \frac{\Pr\left(X \in (-\infty, x] \cap I\right)}{\Pr(X \in I)}.\end{aligned}$$

Differentiating with respect to x, we obtain the conditional PDF of X,

$$f_{X|A}(x) = \frac{f_X(x)}{\Pr(X \in I)}$$

for any $x \in I$. In words, the conditional PDF of X becomes a scaled version of $f_X(\cdot)$ whenever $x \in I$, and it is equal to zero otherwise. Essentially, this is equivalent to re-normalizing the PDF of X over interval I, accounting for the partial information given by $X \in I$.

11.2.1 Conditioning on Values

Suppose that X and Y form a pair of random variables with joint PDF $f_{X,Y}(\cdot,\cdot)$. With great care, it is possible and desirable to define the conditional PDF of X, conditioned on $Y = y$. Special attention must be given to this situation because the event $\{Y = y\}$ has probability zero whenever Y is a continuous random variable. Still, when X and Y are jointly continuous and for any $y \in \mathbb{R}$ such that $f_Y(y) > 0$, we can defined the conditional PDF of X given $Y = y$ as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{11.3}$$

Intuitively, this definition is motivated by the following property. For small Δ_x and Δ_y, we can write

$$\begin{aligned}
&\Pr(x \leq X \leq x + \Delta_x | y \leq Y \leq y + \Delta_y) \\
&= \frac{\Pr(x \leq X \leq x + \Delta_x, y \leq Y \leq y + \Delta_y)}{\Pr(y \leq Y \leq y + \Delta_y)} \\
&\approx \frac{f_{X,Y}(x,y)\Delta_x\Delta_y}{f_Y(y)\Delta_y} = \frac{f_{X,Y}(x,y)}{f_Y(y)}\Delta_x.
\end{aligned}$$

Thus, loosely speaking, $f_{X|Y}(x|y)\Delta_x$ represents the probability that X lies close to x, given that Y is near y.

Using this definition, it is possible to compute the probabilities of events associated with X conditioned on a specific value of random variable Y,

$$\Pr(X \in S|Y = y) = \int_S f_{X|Y}(x|y)dx.$$

A word of caution is in order. The technical difficulty that surfaces when conditioning on $\{Y = y\}$ stems from the fact that $\Pr(Y = y) = 0$. Remember that, in general, the notion of conditional probability is only defined for non-vanishing conditions. Although we were able to circumvent this issue, care must be taken when dealing with the conditional PDF of the form (11.3), as it only provides valuable insight when the random variables (X, Y) are jointly continuous.

Example 97. *Consider the experiment where an outcome (ω_1, ω_2) is selected at random from the unit circle. Let $X = \omega_1$ and $Y = \omega_2$. We wish to compute the conditional PDF of X given that $Y = 0.5$.*

First, we compute the marginal PDF of Y evaluated at $Y = 0.5$,

$$f_Y(0.5) = \int_{\mathbb{R}} f_{X,Y}(x, 0.5) dx = \int_{\frac{\sqrt{3}}{2}}^{\frac{\sqrt{3}}{2}} \frac{1}{\pi} dx = \frac{\sqrt{3}}{\pi}.$$

We then apply definition (11.3) to obtain the desired conditional PDF of X,

$$\begin{aligned} f_{X|Y}(x|0.5) &= \frac{f_{X,Y}(x, 0.5)}{f_Y(0.5)} = \frac{\pi}{\sqrt{3}} f_{X,Y}(x, 0.5) \\ &= \begin{cases} \frac{1}{\sqrt{3}} & |x| \leq \frac{\sqrt{3}}{2} \\ 0 & \textit{otherwise.} \end{cases} \end{aligned}$$

11.2.2 Conditional Expectation

The conditional expectation of a function $g(Y)$ is simply the integral of $g(Y)$ weighted by the proper conditional PDF,

$$\begin{aligned} E[g(Y)|X = x] &= \int_{\mathbb{R}} g(y) f_{Y|X}(y|x) dy \\ E[g(Y)|S] &= \int_{\mathbb{R}} g(y) f_{Y|S}(y) dy. \end{aligned}$$

Note again that the function

$$h(x) = \mathrm{E}[Y|X = x]$$

defines a random variable since the conditional expectation of Y may vary as a function of X. After all, a conditional expectation is itself a random variable.

Example 98. *An analog communication system transmits a random signal over a noisy channel. The transmit signal X and the additive noise N are both* *standard Gaussian random variables, and they are independent. The signal received at the destination is equal to*

$$Y = X + N.$$

We wish to estimate the value of X conditioned on $Y = y$.

For this problem, the joint PDF of X and Y is

$$f_{X,Y}(x,y) = \frac{1}{2\pi} \exp\left(-\frac{2x^2 - 2xy + y^2}{2} \right)$$

and the conditional distribution of X given Y becomes

$$\begin{aligned} f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} &= \frac{\frac{1}{2\pi} \exp\left(-\frac{2x^2-2xy+y^2}{2} \right)}{\frac{1}{2\sqrt{\pi}} \exp\left(-\frac{y^2}{4} \right)} \\ &= \frac{1}{\sqrt{\pi}} \exp\left(-\frac{4x^2 - 4xy + y^2}{4} \right). \end{aligned}$$

By inspection, we recognize that this conditional PDF is a Gaussian distribution with parameters $m = y/2$ and $\sigma^2 = 1/2$. A widespread algorithm employed to perform the desired task is called the minimum mean square error (MMSE) estimator. In the present case, the MMSE estimator reduces to the conditional expectation of X given $Y = y$, which is

$$\mathrm{E}[X|Y = y] = \int_{\mathbb{R}} x f_{X,Y}(x|y) dx = \frac{y}{2}.$$

11.2.3 Derived Distributions

Suppose X_1 and X_2 are jointly continuous random variables. Furthermore, let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$, where $g_1(\cdot,\cdot)$ and $g_2(\cdot,\cdot)$ are real-valued functions. Under certain conditions, the pair of random variables (Y_1, Y_2) will also be continuous. Deriving the joint PDF of (Y_1, Y_2) can get convoluted, a task we forgo. It requires the skillful application of vector calculus. Nevertheless, we examine the case where a simple expression for $f_{Y_1,Y_2}(\cdot,\cdot)$ exists.

Consider the scenario where the functions $g_1(\cdot,\cdot)$ and $g_2(\cdot,\cdot)$ are totally differentiable, with Jacobian determinant

$$\begin{aligned} J(x_1, x_2) &= \det \left[\begin{array}{cc} \frac{\partial g_1}{\partial x_1}(x_1, x_2) & \frac{\partial g_1}{\partial x_2}(x_1, x_2) \\ \frac{\partial g_2}{\partial x_1}(x_1, x_2) & \frac{\partial g_2}{\partial x_2}(x_1, x_2) \end{array} \right] \\ &= \frac{\partial g_1}{\partial x_1}(x_1, x_2) \frac{\partial g_2}{\partial x_2}(x_1, x_2) - \frac{\partial g_1}{\partial x_2}(x_1, x_2) \frac{\partial g_2}{\partial x_1}(x_1, x_2) \neq 0. \end{aligned}$$

Also, assume that the system of equations

$$\begin{aligned} g_1(x_1, x_2) &= y_1 \\ g_2(x_1, x_2) &= y_2 \end{aligned}$$

has a unique solution. We express this solution using $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$. Then, the random variables (Y_1, Y_2) are jointly continuous with joint PDF

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{|J(x_1, x_2)|} \tag{11.4}$$

where $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$. Note the close resemblance between this equation and the derived distribution of (9.4). Looking back at Chapter 9 offers an idea of what proving this result entails. It also hints at how this equation can be modified to accommodate non-unique mappings.

Example 99. *An important application of* (11.4) *pertains to the properties of Gaussian vectors. Suppose that X_1 and X_2 are jointly continuous random variables, and let*

$$\mathbf{X} = \left[\begin{array}{c} X_1 \\ X_2 \end{array} \right].$$

Define the mean of $\mathbf{X}$ by

$$\mathbf{m} = \mathrm{E}[\mathbf{X}] = \left[\begin{array}{c} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \end{array} \right].$$

and its covariance by

$$\begin{aligned} \Sigma &= \mathrm{E}\left[(\mathbf{X} - \mathbf{m}) (\mathbf{X} - \mathbf{m})^T \right] \\ &= \left[\begin{array}{cc} \mathrm{E}\left[(X_1 - m_1)^2 \right] & \mathrm{E}[(X_1 - m_1)(X_2 - m_2)] \\ \mathrm{E}[(X_2 - m_2)(X_1 - m_1)] & \mathrm{E}\left[(X_2 - m_2)^2 \right] \end{array} \right]. \end{aligned}$$

Random variables X_1 and X_2 are said to be jointly Gaussian provided that their joint PDF is of the form

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}) \right).$$

Assume that the random variables Y_1 and Y_2 are generated through the matrix equation

$$\mathbf{Y} = \left[\begin{array}{c} Y_1 \\ Y_2 \end{array} \right] = A\mathbf{X} + \mathbf{b},$$

where A is a 2×2 invertible matrix and $\mathbf{b}$ is a constant vector. In this case, $\mathbf{X} = A^{-1}(\mathbf{Y} - \mathbf{b})$ and the corresponding Jacobian determinant is

$$J(x_1, x_2) = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = |A|.$$

Applying (11.4)*, we gather that the joint PDF of (Y_1, Y_2) is expressed as*

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}|A|} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^T \Sigma^{-1} (\mathbf{x} - \mathbf{m})\right) \\ &= \frac{1}{2\pi|A\Sigma A|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(A^{-1}(\mathbf{y} - \mathbf{b}) - \mathbf{m}\right)^T \Sigma^{-1} \left(A^{-1}(\mathbf{y} - \mathbf{b}) - \mathbf{m}\right)\right) \\ &= \frac{1}{2\pi|A\Sigma A|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{b} - A\mathbf{m})^T (A\Sigma A)^{-1} (\mathbf{y} - \mathbf{b} - A\mathbf{m})\right). \end{aligned}$$

Looking at this equation, we conclude that random variables Y_1 and Y_2 are also jointly Gaussian, as their joint PDF possesses the proper form. It should come as no surprise that the mean of $\mathbf{Y}$ is $\mathrm{E}[\mathbf{Y}] = A\mathbf{m} + \mathbf{b}$ and its covariance matrix is equal to

$$\mathrm{E}\left[(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^T\right] = A\Sigma A.$$

In other words, a non-trivial affine transformation of a two-dimensional Gaussian vector yields another Gaussian vector. This admirable property generalizes to higher dimensions. Indeed, if $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ where A is an $n \times n$ invertible matrix and $\mathbf{X}$ is a Gaussian random vector, then $\mathbf{Y}$ remains a Gaussian random vector. Furthermore, to obtain the derived distribution of the latter vector, it suffices to compute its mean and covariance, and substitute the resulting parameters in the general form of the Gaussian PDF. Collectively, the features of joint Gaussian random vectors underly many contemporary successes of engineering.

11.3 Independence

Two random variables X and Y are mutually *independent* if their joint CDF is the product of their respective CDFs,

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

for $x, y \in \mathbb{R}$. For jointly continuous random variables, this definition necessarily implies that the joint PDF $f_{X,Y}(\cdot, \cdot)$ is the product of their marginal PDFs,

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y) = \frac{dF_X}{dx}(x)\frac{dF_Y}{dy}(y) = f_X(x)f_Y(y)$$

for $x, y \in \mathbb{R}$. Furthermore, we gather from (11.3) that the conditional PDF of Y given $X = x$ is equal to the marginal PDF of Y whenever X and Y are independent,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y)$$

provided of course that $f_X(x) \neq 0$. Additionally, if X and Y are independent random variables, then the two events $\{X \in S\}$ and $\{Y \in T\}$ are also independent,

$$\begin{aligned}\Pr(X \in S, Y \in T) &= \int_S \int_T f_{X,Y}(x, y) dy dx \\ &= \int_S f_X(x) dx \int_T f_Y(y) dy \\ &= \Pr(X \in S) \Pr(Y \in T).\end{aligned}$$

Example 100. *Consider a random experiment where an outcome (ω_1, ω_2) is selected at random from the unit square. Let $X = \omega_1$, $Y = \omega_2$ and $U = \omega_1 + \omega_2$. We wish to show that X and Y are independent, but that X and U are not independent.*

We begin by computing the joint CDF of X and Y. For $x, y \in [0, 1]$, we have

$$F_{X,Y}(x, y) = \int_0^x \int_0^y d\zeta d\xi = xy = F_X(x)F_Y(y).$$

More generally, if $x, y \in \mathbb{R}^2$, we get

$$\begin{aligned}F_{X,Y}(x, y) &= \int_{-\infty}^x \int_{-\infty}^y \mathbf{1}_{[0,1]^2}(\xi, \zeta) d\zeta d\xi \\ &= \int_{-\infty}^x \mathbf{1}_{[0,1]}(\xi) d\xi \int_{-\infty}^y \mathbf{1}_{[0,1]}(\zeta) d\zeta = F_X(x)F_Y(y).\end{aligned}$$

Thus, we gather that X and Y are independent.

Next, we show that X and U are not independent. Note that $F_U(1) = 0.5$ and $F_X(0.5) = 0.5$. Consider the joint CDF of X and U evaluated at $(0.5, 1)$,

$$\begin{aligned} F_{X,U}(0.5, 1) &= \int_0^{\frac{1}{2}} \int_0^{1-\xi} d\zeta d\xi = \int_0^{\frac{1}{2}} (1 - \xi) d\xi \\ &= \frac{3}{8} \neq F_X(0.5) F_U(1). \end{aligned}$$

Clearly, random variables X and U are not independent.

11.3.1 Sums of Continuous Random Variables

As mentioned before, sums of independent random variables are frequently encountered in engineering. We therefore turn to the question of determining the distribution of a sum of independent continuous random variables in terms of the PDFs of its constituents. If X and Y are independent random variables, the distribution of their sum $U = X + Y$ can be obtained by using the *convolution* operator. Let $f_X(\cdot)$ and $f_Y(\cdot)$ be the PDFs of X and Y, respectively. The convolution of $f_X(\cdot)$ and $f_Y(\cdot)$ is the function defined by

$$\begin{aligned} (f_X * f_Y)(u) &= \int_{-\infty}^{\infty} f_X(\xi) f_Y(u - \xi) d\xi \\ &= \int_{-\infty}^{\infty} f_X(u - \zeta) f_Y(\zeta) d\zeta. \end{aligned}$$

The PDF of the sum $U = X + Y$ is the convolution of the individual densities $f_X(\cdot)$ and $f_Y(\cdot)$,

$$f_U(u) = (f_X * f_Y)(u).$$

To show that this is indeed the case, we first consider the CDF of U,

$$\begin{aligned} F_U(u) &= \Pr(U \leq u) = \Pr(X + Y \leq u) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{u-\xi} f_{X,Y}(\xi, \zeta) d\zeta d\xi \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{u-\xi} f_Y(\zeta) d\zeta f_X(\xi) d\xi \\ &= \int_{-\infty}^{\infty} F_Y(u - \xi) f_X(\xi) d\xi. \end{aligned}$$

Taking the derivative of $F_U(u)$ with respect to u, we obtain

$$\begin{aligned}\frac{d}{du}F_U(u) &= \frac{d}{du}\int_{-\infty}^{\infty} F_Y(u-\xi)f_X(\xi)d\xi \\ &= \int_{-\infty}^{\infty} \frac{d}{du}F_Y(u-\xi)f_X(\xi)d\xi \\ &= \int_{-\infty}^{\infty} f_Y(u-\xi)f_X(\xi)d\xi.\end{aligned}$$

Notice the judicious use of the fundamental theorem of calculus. This shows that $f_U(u) = (f_X * f_Y)(u)$.

Example 101 (Sum of Uniform Random Variables)**.** *Suppose that two numbers are independently selected from the interval* $[0,1]$*, each with a uniform distribution. We wish to compute the PDF of their sum. Let* X *and* Y *be random variables describing the two choices, and let* $U = X + Y$ *represent their sum. The PDFs of* X *and* Y *are*

$$f_X(\xi) = f_Y(\xi) = \begin{cases} 1 & 0 \leq \xi \leq 1 \\ 0 & \textit{otherwise.} \end{cases}$$

The PDF of their sum is therefore equal to

$$f_U(u) = \int_{-\infty}^{\infty} f_X(u-\xi)f_Y(\xi)d\xi.$$

Since $f_Y(y) = 1$ *when* $0 \leq y \leq 1$ *and zero otherwise, this integral becomes*

$$f_U(u) = \int_0^1 f_X(u-\xi)d\xi = \int_0^1 \mathbf{1}_{[0,1]}(u-\xi)d\xi.$$

The integrand above is zero unless $0 \leq u - \xi \leq 1$ *(i.e., unless* $u - 1 \leq \xi \leq u$*). Thus, if* $0 \leq u \leq 1$*, we get*

$$f_U(u) = \int_0^u d\xi = u;$$

while, if $1 < u \leq 2$*, we obtain*

$$f_U(u) = \int_{u-1}^1 d\xi = 2 - u.$$

If $u < 0$ or $u > 2$, the value of the PDF becomes zero. Collecting these results, we can write the PDF of U as

$$f_U(u) = \begin{cases} u & 0 \leq u \leq 1, \\ 2 - u & 1 < u \leq 2, \\ 0 & \textit{otherwise.} \end{cases}$$

Example 102 (Sum of Exponential Random Variables)**.** *Two numbers are selected independently from the positive real numbers, each according to an exponential distribution with parameter λ. We wish to find the PDF of their sum. Let X and Y represent these two numbers, and denote this sum by $U = X + Y$. The random variables X and Y have PDFs*

$$f_X(\xi) = f_Y(\xi) = \begin{cases} \lambda e^{-\lambda \xi} & \xi \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

When $u \geq 0$, we can use the convolution formula and write

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_X(u - \xi) f_Y(\xi) d\xi \\ &= \int_0^u \lambda e^{-\lambda(u-\xi)} \lambda e^{-\lambda \xi} d\xi \\ &= \int_0^u \lambda^2 e^{-\lambda u} d\xi = \lambda^2 u e^{-\lambda u}. \end{aligned}$$

On the other hand, if $u < 0$ then we get $f_U(u) = 0$. The PDF of U is given by

$$f_U(u) = \begin{cases} \lambda^2 u e^{-\lambda u} & u \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

This is an Erlang distribution with parameter $m = 2$ and $\lambda > 0$.

Example 103 (Sum of Gaussian Random Variables)**.** *It is an interesting and important fact that the sum of two independent Gaussian random variables is itself a Gaussian random variable. Suppose X is Gaussian with mean m_1 and variance σ_1^2, and similarly Y is Gaussian with mean m_2 and variance σ_2^2, then $U = X + Y$ has a Gaussian density with mean $m_1 + m_2$ and variance $\sigma_1^2 + \sigma_2^2$. We will show this property in the special case where both random variables* *are standard normal random variable. The general case can be attained in a similar manner, but the computations are somewhat tedious.*

Suppose X and Y are two independent Gaussian random variables with PDFs

$$f_X(\xi) = f_Y(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}.$$

Then, the PDF of $U = X + Y$ is given by

$$\begin{aligned} f_U(u) = (f_X * f_Y)(u) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{(u-\xi)^2}{2}} e^{-\frac{\xi^2}{2}} d\xi \\ &= \frac{1}{2\pi} e^{-\frac{u^2}{4}} \int_{-\infty}^{\infty} e^{-\left(\xi - \frac{u}{2}\right)^2} d\xi \\ &= \frac{1}{2\sqrt{\pi}} e^{-\frac{u^2}{4}} \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-\left(\xi - \frac{u}{2}\right)^2} d\xi \right). \end{aligned}$$

The expression within the parentheses is equal to one since the integrant is a Gaussian PDF with $m = u/2$ and $\sigma^2 = 1/2$. Thus, we obtain

$$f_U(u) = \frac{1}{\sqrt{4\pi}} e^{-\frac{u^2}{4}},$$

which verifies that U is indeed Gaussian.

Let X and Y be independent random variables. Consider the random variable $U = X + Y$. The moment generating function of U is given by

$$\begin{aligned} M_U(s) &= \mathrm{E}\left[e^{sU}\right] = \mathrm{E}\left[e^{s(X+Y)}\right] \\ &= \mathrm{E}\left[e^{sX} e^{sY}\right] = \mathrm{E}\left[e^{sX}\right] \mathrm{E}\left[e^{sY}\right] \\ &= M_X(s) M_Y(s). \end{aligned}$$

That is, the moment generating function of the sum of two independent random variables is the product of the individual moment generating functions.

Example 104 (Sum of Gaussian Random Variables)**.** *In this example, we revisit the problem of adding independent Gaussian variables using moment generating functions. Again, let X and Y denote the two independent Gaussian variables. We denote the mean and variance of X by m_1 and σ_1^2. Likewise, we represent the mean and variance of Y by m_2 and σ_2^2. We wish to show that the sum $U = X + Y$ is Gaussian with parameters $m_1 + m_2$ and $\sigma_1^2 + \sigma_2^2$.*

The moment generating functions of X and Y are

$$M_X(s) = e^{m_1 s + \frac{\sigma_1^2 s^2}{2}}$$
$$M_Y(s) = e^{m_2 s + \frac{\sigma_2^2 s^2}{2}}$$

The moment generating function of $U = X + Y$ is therefore equal to

$$M_U(s) = M_X(s) M_Y(s) = \exp\left((m_1 + m_2)s + \frac{(\sigma_1^2 + \sigma_2^2)s^2}{2} \right),$$

which demonstrates that U is a Gaussian random variable with mean $m_1 + m_2$ and variance $\sigma_1^2 + \sigma_2^2$, as anticipated.

Further Reading

1. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 3.5.

2. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Chapter 6, Section 7.5.

3. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Chapter 5, Sections 6.1–6.3.

4. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Chapters 7–9.

Chapter 12

Convergence, Sequences and Limit Theorems

Some of the most astonishing results in probability are related to the properties of sequences of random variables and the convergence of empirical distributions. From an engineering viewpoint, these results are important as they enable the efficient design of complex systems with very small probabilities of failure. Concentration behavior facilitates true economies of scale.

12.1 Types of Convergence

The premise on which most probabilistic convergence results lie is a sequence of random variables $X_1, X_2, \ldots$ and a limiting random variable X, all of which are defined on the same probability space. Recall that a random variable is a function of the outcome of a random experiment. The above statement stipulate that all the random variables listed above are functions of the outcome of a same experiment.

Statements that can be made about a sequence of random variables range from simple assertions to more intricate claims. For instance, the sequence may appear to move toward a deterministic quantity or to behave increasingly akin to a certain function. Alternatively, the CDFs of the random variables in the sequence may appear to approach a precise function. Being able to recognize specific patterns within the sequence is key in establishing converge results. The various statement one can make about the sequence $X_1, X_2, \ldots$ lead to the different types of convergence encountered in probability. Below, we discuss briefly three types of convergence.

Example 105. *Suppose that $X_1, X_2, \ldots$ is a sequence of independent Gaussian random variables, each with mean m and variance σ^2. Define the partial sums*

$$S_n = \sum_{i=1}^{n} X_i, \tag{12.1}$$

and consider the sequence

$$S_1, \frac{S_2}{2}, \frac{S_3}{3}, \ldots \tag{12.2}$$

We know that affine transformations of Gaussian random variables remain Gaussian. Furthermore, we know that sums of jointly Gaussian random variables are also Gaussian. Thus, S_n/n possesses a Gaussian distribution with mean

$$\mathrm{E}\left[\frac{S_n}{n}\right] = \frac{\mathrm{E}[S_n]}{n} = \frac{\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]}{n} = m$$

and variance

$$\mathrm{Var}\left[\frac{S_n}{n}\right] = \frac{\mathrm{Var}[S_n]}{n^2} = \frac{\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]}{n^2} = \frac{\sigma^2}{n}.$$

It appears that the PDF of S_n/n concentrates around m as n approaches infinity. That is, the sequence in (12.2) *seems to become increasingly predictable.*

Example 106. *Again, let $X_1, X_2, \ldots$ be the sequence described above, and let S_n be defined according to* (12.1). *This time, we wish to characterize the properties of*

$$S_1 - m, \frac{S_2 - 2m}{\sqrt{2}}, \frac{S_3 - 3m}{\sqrt{3}}, \ldots$$

From our current discussion, we know that $(S_n - nm)/\sqrt{n}$ is a Gaussian random variables. We can compute its mean and variance as follows

$$\mathrm{E}\left[\frac{S_n - nm}{\sqrt{n}}\right] = \frac{\mathrm{E}[S_n - nm]}{\sqrt{n}} = 0$$

$$\mathrm{Var}\left[\frac{S_n - mn}{\sqrt{n}}\right] = \frac{\mathrm{Var}[S_n - mn]}{n} = \frac{\mathrm{Var}[S_n]}{n} = \sigma^2.$$

No matter how large n is, the random variable $(S_n - nm)/\sqrt{n}$ has a Gaussian distribution with mean zero and variance σ^2. Intriguingly, the distributions remains invariant throughout the sequence.

12.1.1 Convergence in Probability

The basic concept behind the definition of *convergence in probability* is that the probability that a random variable deviates from its typical behavior becomes less likely as the sequence progresses. Formally, a sequence $X_1, X_2, \ldots$ of random variables converges in probability to X if for every $\epsilon > 0$,

$$\lim_{n \to \infty} \Pr\left(|X_n - X| \geq \epsilon\right) = 0.$$

In Example 105, the sequence $\{S_n/n\}$ converges in probability to m.

12.1.2 Mean Square Convergence

We say that a sequence $X_1, X_2, \ldots$ of random variables *converges in mean square* to X if

$$\lim_{n \to \infty} \mathrm{E}\left[|X_n - X|^2\right] = 0.$$

That is, the second moment of the difference between X_n and X vanishes as n goes to infinity. Convergence in the mean square sense implies convergence in probability.

Proposition 9. *Let $X_1, X_2, \ldots$ be a sequence of random variables that converge in mean square to X. Then, the sequence $X_1, X_2, \ldots$ also converges to X in probability.*

Proof. Suppose that $\epsilon > 0$ is fixed. The sequence $X_1, X_2, \ldots$ converges in mean square to X. Thus, for $\delta > 0$, there exists an N such that $n \geq N$ implies

$$\mathrm{E}\left[|X_n - X|^2\right] < \delta.$$

If we apply the Chebyshev inequality to $X_n - X$, we get

$$\Pr\left(|X_n - X| \geq \epsilon\right) \leq \frac{\mathrm{E}\left[|X_n - X|^2\right]}{\epsilon^2} < \frac{\delta}{\epsilon^2}$$

Since δ can be made arbitrarily small, we conclude that this sequence also converges to X in probability. □

12.1.3 Convergence in Distribution

A sequence $X_1, X_2, \ldots$ of random variables is said to *converge in distribution* to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at every point $x \in \mathbb{R}$ where $F_X(\cdot)$ is continuous. This type of convergence is also called *weak convergence.*

Example 107. *Let X_n be a continuous random variable that is uniformly distributed over $[0, 1/n]$. Then, the sequence $X_1, X_2, \ldots$ converges in distribution to* 0.

In this example, $X = 0$ and

$$F_X(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Furthermore, for every $x < 0$, we have $F_{X_n}(x) = 0$; and for every $x > 0$, we have

$$\lim_{n\to\infty} F_{X_n}(x) = 1.$$

Hence, the sequence $X_1, X_2, \ldots$ converges in distribution to a constant.

12.2 The Law of Large Numbers

The law of large numbers focuses on the convergence of empirical averages. Although, there are many versions of this law, we only state its simplest form below. Suppose that $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variable, each with finite second moment. Furthermore, for $n \geq 1$, define the empirical sum

$$S_n = \sum_{i=1}^{n} X_n.$$

The law of large number asserts that the sequence of empirical averages,

$$\frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_n,$$

converges in probability to the mean $\mathrm{E}[X]$.

Theorem 4 (Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean* $\mathrm{E}[X]$ *and finite variance. For every $\epsilon > 0$, we have*

$$\lim_{n \to \infty} \Pr \left(\left| \frac{S_n}{n} - \mathrm{E}[X] \right| \geq \epsilon \right) = \lim_{n \to \infty} \Pr \left(\left| \frac{X_1 + \cdots + X_n}{n} - \mathrm{E}[X] \right| \geq \epsilon \right) = 0.$$

Proof. Taking the expectation of the empirical average, we obtain

$$\mathrm{E} \left[\frac{S_n}{n} \right] = \frac{\mathrm{E}\left[S_n\right]}{n} = \frac{\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]}{n} = \mathrm{E}[X].$$

Using independence, we also have

$$\mathrm{Var} \left[\frac{S_n}{n} \right] = \frac{\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]}{n^2} = \frac{\mathrm{Var}[X]}{n}.$$

As n goes to infinity, the variance of the empirical average S_n/n vanishes. Thus, we showed that the sequence $\{S_n/n\}$ of empirical averages converges in mean square to $\mathrm{E}[X]$ since

$$\lim_{n \to \infty} \mathrm{E} \left[\left| \frac{S_n}{n} - \mathrm{E}[X] \right|^2 \right] = \lim_{n \to \infty} \mathrm{Var} \left[\frac{S_n}{n} \right] = 0.$$

To get convergence in probability, we apply the Chebyshev inequality, as we did in Proposition 9,

$$\Pr \left(\left| \frac{S_n}{n} - \mathrm{E}[X] \right| \geq \epsilon \right) \leq \frac{\mathrm{Var}\left[\frac{S_n}{n}\right]}{\epsilon^2} = \frac{\mathrm{Var}[X]}{n \epsilon^2},$$

which clearly goes to zero as n approaches infinity. □

Example 108. *Suppose that a die is thrown repetitively. We are interested in the average number of times a six shows up on the top face, as the number of throws becomes very large.*

Let D_n be a random variable that represent the number on the nth roll. Also, define the random variable $X_n = \mathbf{1}_{\{D_n = 6\}}$. Then, X_n is a Bernoulli random variable with parameter $p = 1/6$, and the empirical average S_n/n is equal to the number of times a six is observed divided by the total number of rolls. By the law of large numbers, we have

$$\lim_{n \to \infty} \Pr \left(\left| \frac{S_n}{n} - \frac{1}{6} \right| \geq \epsilon \right) = 0.$$

That is, as the number of rolls increases, the average number of times a six is observe converges to the probability of getting a six.

12.2.1 Heavy-Tailed Distributions*

There are situations where the law of large numbers does not apply. For example, when dealing with *heavy-tailed distributions*, one needs to be very careful. In this section, we study Cauchy random variables in greater details. First, we show that the sum of two independent Cauchy random variables is itself a Cauchy random variable.

Let X_1 and X_2 be two independent Cauchy random variables with parameter γ_1 and γ_2, respectively. We wish to compute the PDF of $S = X_1 + X_2$. For continuous random variable, the PDF of S is given by the convolution of $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$. Thus, we can write

$$\begin{aligned} f_S(x) &= \int_{-\infty}^{\infty} f_{X_1}(\xi) f_{X_2}(x - \xi) d\xi \\ &= \int_{-\infty}^{\infty} \frac{\gamma_1}{\pi \left(\gamma_1^2 + \xi^2\right)} \frac{\gamma_2}{\pi \left(\gamma_2^2 + (x - \xi)^2\right)} d\xi. \end{aligned}$$

This integral is somewhat difficult to solve. We therefore resort to complex analysis and contour integration to get a solution. Let C be a contour that goes along the real line from $-a$ to a, and then counterclockwise along a semicircle centered at zero. For a large enough, Cauchy's residue theorem requires that

$$\begin{aligned} \oint_C f_{X_1}(z) f_{X_2}(x - z) dz &= \oint_C \frac{\gamma_1}{\pi \left(\gamma_1^2 + z^2\right)} \frac{\gamma_2}{\pi \left(\gamma_2^2 + (x - z)^2\right)} dz \\ &= 2\pi i \left(\operatorname{Res}(g, i\gamma_1) + \operatorname{Res}(g, x + i\gamma_2) \right) \end{aligned} \tag{12.3}$$

where we have implicitly defined the function

$$\begin{aligned} g(z) &= \frac{\gamma_1 \gamma_2}{\pi^2 \left(\gamma_1^2 + z^2\right) \left(\gamma_2^2 + (z - x)^2\right)} \\ &= \frac{\gamma_1 \gamma_2}{\pi^2 (z - i\gamma_1)(z + i\gamma_1)(z - x + i\gamma_2)(z - x - i\gamma_2)}. \end{aligned}$$

Only two residues are contained within the enclosed region. Because they are simple poles, their values are given by

$$\begin{aligned} \operatorname{Res}(g, i\gamma_1) &= \lim_{z \to i\gamma_1} (z - i\gamma_1) g(z) = \frac{\gamma_2}{2i\pi^2 \left((x - i\gamma_1)^2 + \gamma_2^2\right)} \\ \operatorname{Res}(g, x + i\gamma_2) &= \lim_{z \to x + i\gamma_2} (z - x - i\gamma_2) g(z) = \frac{\gamma_1}{2i\pi^2 \left((x + i\gamma_2)^2 + \gamma_1^2\right)}. \end{aligned}$$

It follows that

$$\begin{aligned}
&2\pi i \left(\mathrm{Res}(g, i\gamma_1) + \mathrm{Res}(g, x + i\gamma_2) \right) \\
&= \frac{\gamma_2}{\pi \left((x - i\gamma_1)^2 + \gamma_2^2 \right)} + \frac{\gamma_1}{\pi \left((x + i\gamma_2)^2 + \gamma_1^2 \right)} \\
&= \frac{(\gamma_1 + \gamma_2)}{\pi \left((\gamma_1 + \gamma_2)^2 + x^2 \right)}
\end{aligned}$$

The contribution of the arc in (12.3) vanishes as $a \to \infty$. We then conclude that the PDF of S is equal to

$$f_S(x) = \frac{(\gamma_1 + \gamma_2)}{\pi \left((\gamma_1 + \gamma_2)^2 + x^2 \right)}.$$

The sum of two independent Cauchy random variables with parameters γ_1 and γ is itself a Cauchy random variable with parameter $\gamma_1 + \gamma_2$.

Let $X_1, X_2, \ldots$ form a sequence of independent Cauchy random variables, each with parameter γ. Also, consider the empirical sum

$$S_n = \sum_{i=1}^{n} X_n.$$

Using mathematical induction and the aforementioned fact, it is possible to show that S_n is a Cauchy random variable with parameter $n\gamma$. Furthermore, for $x \in \mathbb{R}$, the PDF of the empirical average S_n/n is given by

$$\frac{f_{S_n}(nx)}{\left| \frac{1}{n} \right|} = \frac{n^2 \gamma}{\pi \left(n^2 \gamma^2 + (nx)^2 \right)} = \frac{\gamma}{\pi \left(\gamma^2 + x^2 \right)},$$

where we have used the methodology developed in Chapter 9. Amazingly, the empirical average of a sequence of independent Cauchy random variables, each with parameter γ, remains a Cauchy random variable with the same parameter. Clearly, the law of large numbers does not apply to this scenario. Note that our version of the law of large numbers requires random variables to have finite second moments, a condition that is clearly violated by the Cauchy distribution. This explains why convergence does not take place in this situation.

12.3 The Central Limit Theorem

The *central limit theorem* is a remarkable result in probability; it partly explains the prevalence of Gaussian random variables. In some sense, it captures the behavior of large sums of small, independent random components.

Theorem 5 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables, each with mean* $\mathrm{E}[X]$ *and variance σ^2. The distribution of*

$$\frac{S_n - n\mathrm{E}[X]}{\sigma\sqrt{n}}$$

converges in distribution to a standard normal random variable as $n \to \infty$. In particular, for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left(\frac{S_n - n\mathrm{E}[X]}{\sigma\sqrt{n}} \leq x \right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} d\xi.$$

Proof. Initially, we assume that $\mathrm{E}[X] = 0$ and $\sigma^2 = 1$. Furthermore, we only study the situation where the moment generating function of X exists and is finite. Consider the log-moment generating function of X,

$$\Lambda_X(s) = \log M_X(s) = \log \mathrm{E}\left[e^{sX}\right].$$

The first two derivatives of $\Lambda_X(s)$ are equal to

$$\begin{aligned} \frac{d\Lambda_X}{ds}(s) &= \frac{1}{M_X(s)} \frac{dM_X}{ds}(s) \\ \frac{d^2\Lambda_X}{ds^2}(s) &= -\left(\frac{1}{M_X(s)}\right)^2 \left(\frac{dM_X}{ds}(s)\right)^2 + \frac{1}{M_X(s)} \frac{d^2 M_X}{ds^2}(s). \end{aligned}$$

Collecting these results and evaluating the functions at zero, we get $\Lambda_X(0) = 0$, $\frac{d\Lambda_X}{ds}(0) = \mathrm{E}[X] = 0$, and $\frac{d^2\Lambda_X}{ds^2}(0) = \mathrm{E}\left[X^2\right] = 1$. Next, we study the log-moment generating function of $S_n/\sqrt{n}$. Recall that the expectation of a product of independent random variables is equal to the product of their individual expectations. Using this property, we get

$$\begin{aligned} \log \mathrm{E}\left[e^{sS_n/\sqrt{n}}\right] &= \log \mathrm{E}\left[e^{sX_1/\sqrt{n}} \cdots e^{sX_n/\sqrt{n}}\right] \\ &= \log\left(M_X\left(\frac{s}{\sqrt{n}}\right) \cdots M_X\left(\frac{s}{\sqrt{n}}\right)\right) \\ &= n\Lambda_X\left(sn^{-1/2}\right). \end{aligned}$$

To explore the asymptotic behavior of the sequence $\{S_n/\sqrt{n}\}$, we take the limit of $n\Lambda_X\left(sn^{-1/2}\right)$ as $n \to \infty$. In doing so, notice the double use of L'Hôpital's rule,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n^{-1}}\Lambda\left(sn^{-1/2}\right) &= \lim_{n\to\infty}\frac{1}{n^{-2}}\frac{d\Lambda}{ds}\left(sn^{-1/2}\right)\frac{sn^{-3/2}}{2}\\ &= \lim_{n\to\infty}\frac{s}{2n^{-1/2}}\frac{d\Lambda}{ds}\left(sn^{-1/2}\right)\\ &= \lim_{n\to\infty}\frac{s}{n^{-3/2}}\frac{d^2\Lambda}{ds^2}\left(sn^{-1/2}\right)\frac{sn^{-3/2}}{2}\\ &= \lim_{n\to\infty}\frac{s^2}{2}\frac{d^2\Lambda}{ds^2}\left(sn^{-1/2}\right) = \frac{s^2}{2}.\end{aligned}$$

That is, the moment generating function of $S_n/\sqrt{n}$ converges point-wise to $e^{s^2/2}$ as $n \to \infty$. In fact, this implies that

$$\Pr\left(\frac{S_n}{\sqrt{n}} \leq x\right) \to \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{\xi^2}{2}}d\xi$$

as $n \to \infty$. In words, $S_n/\sqrt{n}$ converges in distribution to a standard normal random variable. The more general case where $\mathrm{E}[X]$ and $\mathrm{Var}[X]$ are arbitrary constants can be established in an analog manner by proper scaling of the random variables $X_1, X_2, \ldots$ □

In the last step of the proof, we stated that point-wise convergence of the moment generating functions implies convergence in distribution. This is a sophisticated result that we quote without proof.

12.3.1 Normal Approximation

The central limit theorem can be employed to approximate the CDF of large sums. Again, let

$$S_n = X_1 + \cdots + X_n$$

where $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean $\mathrm{E}[X]$ and variance σ^2. When n is large, the CDF of S_n can be estimated by approximating

$$\frac{S_n - n\mathrm{E}[X]}{\sigma\sqrt{n}}$$

as a standard normal random variable. More specifically, we have

$$\begin{aligned}F_{S_n}(x) = \Pr(S_n \leq x) &= \Pr\left(\frac{S_n - n\mathrm{E}[X]}{\sigma\sqrt{n}} \leq \frac{x - n\mathrm{E}[X]}{\sigma\sqrt{n}}\right)\\ &\approx \Phi\left(\frac{x - n\mathrm{E}[X]}{\sigma\sqrt{n}}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Chapter 8.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Sections 7.2–7.4.

3. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Chapter 7.