MATH 407, PROBABILITY THEORY, FALL 2020

STEVEN HEILMAN

Abstract. These notes closely follow the book of Bertsekas and Tsitsiklis, available here.

Contents

Date: October 13, 2020

1. Introduction

Probability generally asks, "How likely is something going to happen?" In your previous experiences with probability, you most likely dealt with dice rolls or decks of cards. For example, you could ask, "If I roll two fair dice, what is the chance their sum 6?" or "If I roll three fair dice, with what probability will their sum be less than 5?" However, modern probability also concerns events that occur on a continuum. Using an example borrowed from Professor David Aldous, consider a dart that is thrown at a dartboard. Theoretically, there are an infinite number of places on the board that the dart could hit. And we can still ask probabilistic questions such as:

- "With what probability will I be able to hit the bullseye?"
- "With what probability will I miss the board entirely?"
- "Given that I hit the board, with what probability will I hit the bullseye?"
- "What score can I expect to get after five dart throws?" (What is the expected value of the score?)
- "How close will I generally be from my expected score?" (What is the variance of the score?)

Our presentation of the elements of probability theory will allow us to consider both discrete objects (as in dice games or card games) and continuous objects (as in darts thrown at a dartboard). We will begin by giving a precise mathematical definition of a probability in Section 2. We will then discuss the analysis of random numbers, otherwise known as random variables, in Section 3. Since discrete random variables are easier to understand, we will begin with them. The expected value and variance in Section 4 are the most fundamental quantities to compute for random variables. For example, if students in this class take a test and their test scores resemble a "bell curve" or Gaussian, then the expected value, or mean, would be the "center" of the bell curve, and the variance would measure how "wide" the bell curve is. (The standard deviation is the square root of the variance.) We will conclude the course with a detailed discussion of continuous random variables in Section 5.

2. Sets and Probabilities

2.1. **Sets.** In probability theory, a set represents some possible outcomes of some random process. For example, the set $\{1,2,3,4,5,6\}$ has six elements which represent all possible rolls for a six-sided die. The set $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$ has $6 \cdot 6 = 36$ elements, representing all possible ordered dice rolls for two six-sided dice. For example, the ordered pair $(2,3)$ represents a roll of 2 on the first die, and a 3 on the second die. The set $[0,1] \times [0,1]$

in the plane $\mathbb{R}^2$ could represent the set of all possible locations of a dart thrown at a square dartboard.

Eventually, we will assign probabilities to all elements of the set, but for now we will just focus on the sets themselves.

Definition 2.1 (**Set, Element**)**.** A **set** is a collection of objects. Each such object in the set is called an **element** of the set. If A is a set and x is an element of the set A, we write $x \in A$. If x is not an element of A, we write $x \notin A$. The set consisting of no elements is called the **emptyset**, and this set is denoted by $\emptyset$.

Definition 2.2 (**Finite, Countably Infinite**)**.** Let A be a set. We say that the set A is **finite** if there exists a nonnegative integer n such that A can be enumerated as a set of n elements. That is, we can write $A = \{x_1, x_2, \ldots, x_n\}$. We say that the set A is **countably infinite** if A can be enumerated by the positive integers. That is, we can write $A = \{x_1, x_2, x_3, \ldots\}$. We say that the set A is **uncountable** if: A is not finite, and A is not countably infinite.

Example 2.3. The set $\{1, 2, 3, 4, 5, 6\}$ is finite. The set $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ is finite. The set of positive integers $\{1, 2, 3, 4, \ldots\}$ is countably infinite. The set of even positive integers $\{2, 4, 6, 8, \ldots\}$ is countably infinite. We could write the positive even integers in the following way.

$$\{2, 4, 6, 8, \ldots\} = \{k \in \mathbb{R} \colon k/2 \text{ is a positive integer}\}.$$

The last expression is read as "The set of all k in the set of real numbers such that $k/2$ is a positive integer."

The closed interval $[0, 1] = \{x \in \mathbb{R} \colon 0 \leq x \leq 1\}$ is uncountable; this (perhaps counterintuitive) fact is sometimes proven in Real Analysis, Math 131A. That is, there is no way to write $[0, 1]$ as a list $\{x_1, x_2, x_3, \ldots\}$ where $x_i \in [0, 1]$ for every positive integer i.

Definition 2.4 (**Subset**)**.** Let A and B be sets. If every element of A is also an element of B, we say that A is a **subset** of B, and we write $A \subseteq B$, or $B \supseteq A$. If $B \subseteq A$ and $A \subseteq B$, we say that A and B are **equal** and we write $A = B$.

Definition 2.5 (**Universal Set**)**.** In a specific problem, we assume the existence of a sample space, or **universal set** Ω which contains all other sets. The universal set represents all possible outcomes of some random process. We sometimes call the universal set the **universe**. The universe is always assumed to be nonempty.

Example 2.6. We represent the roll of a single six-sided die by the universal set $\Omega = \{1, 2, 3, 4, 5, 6\}$. The set $A = \{1, 2, 3\}$ satisfies $A \subseteq \Omega$.

We can think of throwing darts at a flat, infinite board, so that the universal set is $\Omega = \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) \colon x \in \mathbb{R} \text{ and } y \in \mathbb{R}\}$. We could imagine the dartboard itself as a square subset $[0, 1] \times [0, 1] \subseteq \Omega$. Or, perhaps we could imagine a circular dartboard as a subset $\{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 1\} \subseteq \Omega$.

Definition 2.7 (**Complement**)**.** Suppose A is a subset of some universal set Ω. The **complement** of A in Ω, denoted by A^c, is the set $\{x \in \Omega \colon x \notin A\}$.

Example 2.8. If $\Omega = \{1, 2, 3, 4, 5, 6\}$ and if $A = \{1, 2, 3\}$, then $A^c = \{4, 5, 6\}$.

Note that we always have $\emptyset^c = \Omega$ and $\Omega^c = \emptyset$.

Definition 2.9 (**Union, Intersection**)**.** Let A, B be sets in some universe Ω. The **union** of A and B, denoted $A \cup B$, is the set of elements that are in either A or B. That is,

$$A \cup B = \{x \in \Omega \colon x \in A \text{ or } x \in B\}.$$

The **intersection** of A and B, denoted $A \cap B$, is the set of elements that are in both A and B. That is,

$$A \cap B = \{x \in \Omega \colon x \in A \text{ and } x \in B\}.$$

The **set difference** of A and B, denoted $A \setminus B$, is the set of elements that are in A but not in B. So,

$$A \setminus B = \{x \in A \colon x \notin B\}.$$

Let n be a positive integer. Let $A_1, A_2, \ldots, A_n$ be sets in Ω. We denote

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n.$$

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n.$$

Example 2.10. If $\Omega = \{1, 2, 3, 4, 5, 6\}$, if $A = \{1, 2, 3\}$, and if $B = \{3, 4\}$, then $A \cup B = \{1, 2, 3, 4\}$ and $A \cap B = \{3\}$.

Definition 2.11 (**Countable Union, Countable Intersection**)**.** Let $A_1, A_2, \ldots$ be sets in some universe Ω. The **countable union** of $A_1, A_2, \ldots$, denoted $\cup_{i=1}^{\infty} A_i$ is defined as follows.

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega \colon \exists \text{ a positive integer } j \text{ such that } x \in A_j\}.$$

The **countable intersection** of $A_1, A_2, \ldots$, denoted $\cap_{i=1}^{\infty} A_i$ is defined as follows.

$$\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega \colon x \in A_j, \ \forall \text{ positive integers } j\}.$$

Exercise 2.12. Prove that the set of real numbers $\mathbb{R}$ can be written as the countable union

$$\mathbb{R} = \bigcup_{j=1}^{\infty} [-j, j].$$

(Hint: you should show that the left side contains the right side, and also show that the right side contains the left side.)

Prove that the singleton set $\{0\}$ can be written as

$$\{0\} = \bigcap_{j=1}^{\infty} [-1/j, 1/j].$$

Definition 2.13 (**Disjointness**)**.** Let n be a positive integer. Let A, B be sets in some universe Ω. We say that A and B are **disjoint** if $A \cap B = \emptyset$. A collection of sets $A_1, A_2, \ldots, A_n$ in Ω is said to be a **partition** of Ω if $\cup_{i=1}^{n} A_i = \Omega$, and if, for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$, we have $A_i \cap A_j = \emptyset$.

Remark 2.14. Two or three sets can be visualized with a Venn diagram, though the Venn diagram is no longer very helpful when considering more than three sets.

Exercise 2.15. Let $\Omega = \{1, 2, 3, \ldots, 10\}$. Find sets $A_1, A_2, A_3 \subseteq \Omega$ such that: $A_1 \cap A_2 = \{2, 3\}$, $A_1 \cap A_3 = \{3, 4\}$, $A_2 \cap A_3 = \{3, 5\}$, $A_1 \cap A_2 \cap A_3 = \{3\}$, and such that $A_1 \cup A_2 \cup A_3 = \{2, 3, 4, 5\}$.

The following properties follow from the above definitions.

Proposition 2.16. *Let A, B, C be sets in a universe Ω.*

(i) $A \cup B = B \cup A$.
(ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
(iii) $(A^c)^c = A$.
(iv) $A \cup \Omega = \Omega$.
(v) $A \cup (B \cup C) = (A \cup B) \cup C$.
(vi) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
(vii) $A \cap A^c = \emptyset$.
(viii) $A \cap \Omega = A$.

Exercise 2.17. Using the definitions of intersection, union and complement, prove properties (ii) and (iii). (Hint: to prove property (ii), it may be helpful to first draw a Venn diagram of A, B, C. Now, let $x \in \Omega$. Consider where x could possibly be with respect to A, B, C. For example, we could have $x \in A$, $x \notin B$, $x \in C$. We could also have $x \in A$, $x \in B$, $x \notin C$. And so on. In total, there should be $2^3 = 8$ possibilities for the location of x, with respect to A, B, C. Construct a truth table which considers all eight such possibilities for each side of the purported equality $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.)

Proposition 2.18 (De Morgan's Laws). *Let $A_1, A_2, \ldots$ be sets in some universe Ω. Then*

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c, \qquad \left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c.$$

Proof. We prove the first equality, since the second follows similarly. Suppose $x \in (\cup_{i=1}^{\infty} A_i)^c$. That is, $x \notin \cup_{i=1}^{\infty} A_i$. Recall that $\cup_{i=1}^{\infty} A_i = \{x \in \Omega \colon \exists \text{ a positive integer } j \text{ such that } x \in A_j\}$. Since x is not in the set $\cup_{i=1}^{\infty} A_i$, the negation of the definition of $\cup_{i=1}^{\infty} A_i$ applies to x. That is, x satisfies the negation of the statement: "$\exists$ a positive integer j such that $x \in A_j$". The negation of this statement is: "$\forall$ positive integers j, we have $x \notin A_j$." That is, $\forall$ positive integers j, we have $x \in A_j^c$. By the definition of countable intersection, we conclude that $x \in \cap_{i=1}^{\infty} A_i^c$.

So, we showed that $(\cup_{i=1}^{\infty} A_i)^c \subseteq \cap_{i=1}^{\infty} A_i^c$. To conclude, we must show that $(\cup_{i=1}^{\infty} A_i)^c \supseteq \cap_{i=1}^{\infty} A_i^c$. So, let $x \in \cap_{i=1}^{\infty} A_i^c$. By reversing the above implications, we conclude that $x \in (\cup_{i=1}^{\infty} A_i)^c$. That is, $(\cup_{i=1}^{\infty} A_i)^c \supseteq \cap_{i=1}^{\infty} A_i^c$, and the proof is complete. □

Exercise 2.19. Prove that $(\bigcap_{i=1}^{\infty} A_i)^c = \bigcup_{i=1}^{\infty} A_i^c$.

Exercise 2.20. Let $A_1, A_2, \ldots$ be sets in some universe Ω. Let $B \subseteq \Omega$. Show the following generalization of Proposition 2.16(ii).

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

Exercise 2.21. Let $f \colon \mathbb{R} \to \mathbb{R}$ be a function. Show that

$$\cup_{y \in \mathbb{R}} \{x \in \mathbb{R} \colon f(x) = y\} = \mathbb{R}.$$

Also, show that the union on the left is disjoint. That is, if $y_1 \neq y_2$ and $y_1, y_2 \in \mathbb{R}$, then $\{x \in \mathbb{R} \colon f(x) = y_1\} \cap \{x \in \mathbb{R} \colon f(x) = y_2\} = \emptyset$.

2.2. **Probabilistic Models.**

Definition 2.22. A **probabilistic model** consists of

- A universal set Ω, which represents all possible outcomes of some random process.
- A **probability law P**. Given a set $A \subseteq \Omega$, the probability law assigns a number $\mathbf{P}(A)$ to the set A. A set $A \subseteq \Omega$ is also called an **event**. The number $\mathbf{P}(A)$ denotes the probability that the event A will occur. The probability law satisfies the axioms below.

Axioms for a Probability Law:

(i) For any $A \subseteq \Omega$, we have $\mathbf{P}(A) \geq 0$. (**Nonnegativity**)

(ii) For any $A, B \subseteq \Omega$ such that $A \cap B = \emptyset$, we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

If $A_1, A_2, \ldots \subseteq \Omega$ and $A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \neq j$, then

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k). \qquad (\textbf{Additivity})$$

(iii) We have $\mathbf{P}(\Omega) = 1$. (**Normalization**)

Proposition 2.23. *Let* $\mathbf{P}$ *be a probability law on a universe* Ω. *Let* $A \subseteq \Omega$. *Then* $\mathbf{P}(A) + \mathbf{P}(A^c) = 1$, *and* $\mathbf{P}(A) \in [0, 1]$.

Proof. Let $A \subseteq \Omega$. Then $\Omega = A \cup (A^c)$, and $A \cap (A^c) = \emptyset$ by Proposition 2.16(vii). So, using Axiom (iii) and then Axiom (ii), we have

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(A \cup (A^c)) = \mathbf{P}(A) + \mathbf{P}(A^c).$$

That is, $\mathbf{P}(A) = 1 - \mathbf{P}(A^c)$. Since $\mathbf{P}(A^c) \geq 0$ by Axiom (i), we conclude that $\mathbf{P}(A) \leq 1$. Using Axiom (i) again, we have $\mathbf{P}(A) \geq 0$. In conclusion, $\mathbf{P}(A) \in [0, 1]$. □

Remark 2.24. Since $\mathbf{P}(A) + \mathbf{P}(A^c) = 1$, choosing $A = \emptyset$ shows that $\mathbf{P}(\emptyset) + \mathbf{P}(\Omega) = 1$, so that $\mathbf{P}(\emptyset) = 0$ by Axiom (iii). Consequently, suppose n is a positive integer, and let $A_1, \ldots, A_n \subseteq \Omega$ with $A_i \cap A_j = \emptyset$ whenever $i, j \in \{1, \ldots, n\}$ and $i \neq j$. For any $i > n$, let $A_i = \emptyset$. Then Axiom (ii) implies that

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k) = \sum_{k=1}^{n} \mathbf{P}(A_k).$$

This identity also follows from the first part of Axiom (ii) and by induction on n.

Remark 2.25. If $A, B \subseteq \Omega$ are not disjoint, i.e. if $A \cap B \neq \emptyset$, then it is possible that $\mathbf{P}(A \cup B) \neq \mathbf{P}(A) + \mathbf{P}(B)$.

Example 2.26. Let's return to the example of rolling a single six-sided die. Recall that we used $\Omega = \{1, 2, 3, 4, 5, 6\}$. For a fair die, we define $\mathbf{P}$ so that, for any event $A \subseteq \Omega$, $\mathbf{P}(A)$ is the number of elements of A divided by 6. In particular, $\mathbf{P}(\{i\}) = 1/6$ for each $i \in \Omega$. That is, the probability of rolling any number is $1/6$. Then $\mathbf{P}$ satisfies all axioms of a probability law. (Verify this as an exercise.)

If we think of the axioms as intuitive statements about probabilities, these statements seem to be sensible. Axiom (iii) says that all results together have probability 1. As we have shown in Proposition 2.23, all three axioms show that the probability of any event is some number in the closed interval $[0, 1]$. Axiom (ii) says that if two events have nothing to do with each other, then their probabilities add. For example, the probability of rolling 2 or 3 is equal to the probability of rolling a 2, plus the probability of rolling a 3. Or, written more concisely, $\mathbf{P}(\{2, 3\}) = \mathbf{P}(\{2\}) + \mathbf{P}(\{3\})$.

Note that we can verify Remark 2.25, since

$$\mathbf{P}(\{1, 2\} \cup \{2, 3\}) = \mathbf{P}(\{1, 2, 3\}) = 1/2 \neq 2/3 = \mathbf{P}(\{1, 2\}) + \mathbf{P}(\{2, 3\}).$$

There are many different probability laws $\mathbf{P}$ that could be assigned to the universe $\Omega = \{1, 2, 3, 4, 5, 6\}$. For example, consider an unfair die defined so that, for any $A \subseteq \Omega$, we have

$$\mathbf{P}(A) = \begin{cases} 1, & \text{if } 6 \in A \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathbf{P}$ satisfies all axioms of a probability law. (Verify this as an exercise.) Our interpretation of this probability law $\mathbf{P}$ is that the die will always roll a 6, since $\mathbf{P}(\{6\}) = 1$.

We can generalize the first part of Example 2.26 as follows.

Exercise 2.27 (**Discrete Uniform Probability Law**)**.** Let n be a positive integer. Suppose we are given a finite universe Ω with exactly n elements. Let $A \subseteq \Omega$. Define $\mathbf{P}(A)$ such that $\mathbf{P}(A)$ is the number of elements of A, divided by n. Verify that $\mathbf{P}$ is a probability law. This probability law is referred to as the uniform probability law on Ω, since each element of Ω has the same probability.

More generally, we can compute probabilities for any probability law in any finite universe in the following way.

Example 2.28 (**Discrete Probability Law**)**.** Suppose we are given a finite universe Ω and a probability law $\mathbf{P}$. That is, we are given a discrete probability law. Then for any event $A \subseteq \Omega$, there exists a nonnegative integer n and there exist distinct $a_1, \ldots, a_n \in \Omega$ such that $A = \{a_1, \ldots, a_n\}$. Writing $A = \cup_{k=1}^{n} \{a_i\}$ and applying Remark 2.24, we conclude that

$$\mathbf{P}(A) = \mathbf{P}(\{a_1, \ldots, a_n\}) = \sum_{k=1}^{n} \mathbf{P}(\{a_k\}).$$

Example 2.29. Let's return to our example of rolling two fair six-sided dice. Recall that we use $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ as our sample space. Then, given any $A \subseteq \Omega$, we let $\mathbf{P}$ be the uniform probability law on Ω. It follows from Exercise 2.27 that $\mathbf{P}$ is a probability law on Ω.

Let's compute a few probabilities. Recall that Ω has $6 \cdot 6 = 36$ elements. By the definition of $\mathbf{P}$, we have $\mathbf{P}(1, 1) = 1/36$. In fact, by the definition of $\mathbf{P}$, for any fixed $i, j \in \{1, 2, 3, 4, 5, 6\}$, we have $\mathbf{P}(i, j) = 1/36$.

Let A be the event that both dice rolls are equal. What is $\mathbf{P}(A)$? There are only six dice rolls where both dice are equal. They are: $(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)$. So, by the definition of $\mathbf{P}$, we have $\mathbf{P}(A) = 6/36 = 1/6$.

From Axiom (ii), $\mathbf{P}((1,2) \cup (2,1)) = 2/36$. That is, the probability that the sum of the dice is 3 is 2/36. More generally, let $s \in \{2, 3, 4, \ldots, 12\}$. What is the probability that the sum of the dice is s? To answer this question, let's count the number of ordered pairs $(i, j) \in \Omega$ such that $i + j = s$. If $s \leq 7$, these ordered pairs are $(1, s-1)$, $(2, s-2)$, ..., $(s-1, 1)$. So, there are $s-1$ such ordered pairs. If $s > 7$, these ordered pairs are $(6, s-6)$, $(5, s-5)$, ..., $(s-6, 6)$. So, there are $6 - (s-6) + 1 = 13 - s$ such ordered pairs. Writing $\min(s-1, 13-s)$ for the minimum of the numbers $s-1$ and $13-s$, we conclude there are $\min(s-1, 13-s)$ ordered pairs $(i, j) \in \Omega$ with $i + j = s$. So, the probability that the sum of the dice is s is $(1/36)\min(s-1, 13-s)$ when $2 \leq s \leq 12$.

2.2.1. *Continuous Models.* As we discussed in the Introduction, many interesting probability space are not discrete. Here are some examples of continuous, i.e. non-discrete, probability laws.

Example 2.30. Let $\Omega = [0, 1]$. For any interval of the form $[a, b]$ with $0 \leq a < b \leq 1$, define $\mathbf{P}([a, b]) = b - a$. Then $\mathbf{P}$ defines a probability law on Ω. (Note that there are many more subsets A of Ω for which we would need to define $\mathbf{P}(A)$, but doing so is a complicated endeavor which is outside the scope of this course. This topic, known as measure theory, is covered graduate real analysis and probability classes.)

We can think of $\mathbf{P}$ as a uniform probability law on Ω, since any interval has a probability which is equal to its length. That is, the probability of any interval does not depend on its position. We can perhaps think of $\mathbf{P}$ as giving the height of a random plant, or the length of a random river. However, there is an important difference between the present example and the discrete uniform probability law. In the present case, the probability of any element of Ω is 0. For example, $\mathbf{P}(\{0.37\}) = 0$.

Exercise 2.31. Let $\Omega = \mathbb{R}^2$. Let $A \subseteq \Omega$. Define a probability law $\mathbf{P}$ on Ω so that

$$\mathbf{P}(A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dx dy.$$

We can think of $\mathbf{P}$ as defining the (random) position of a dart, thrown at an infinite dart board. That is, if $A \subseteq \Omega$, then $\mathbf{P}(A)$ is the probability that the dart will land in the set A.

Very that Axiom (iii) holds for $\mathbf{P}$. That is, verify that $\mathbf{P}(\Omega) = 1$. Then, compute the probability that a dart hits a circular board A, where $A = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 1\}$.

Exercise 2.32. Let A, B be subsets of a sample space Ω. Show the following things:

- $A = (A \smallsetminus B) \cup (A \cap B)$, and $(A \smallsetminus B) \cap (A \cap B) = \emptyset$.
- $A \cup B = (A \smallsetminus B) \cup (B \smallsetminus A) \cup (A \cap B)$, and the three sets $(A \smallsetminus B), (B \smallsetminus A), (A \cap B)$ are all disjoint. That is, any two of these sets are disjoint.

Proposition 2.33 (**Properties of Probability Laws**). *Let Ω be a sample space and let $\mathbf{P}$ be a probability law on Ω. Let $A, B, C \subseteq \Omega$.*

- *If $A \subseteq B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.*
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.
- $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.
- $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

Let n be a positive integer. Let $A_1, \ldots, A_n \subseteq \Omega$. Then

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \mathbf{P}(A_k).$$

Proof. Let $A \subseteq B$. Then $B = (B \cap A) \cup (B \cap A^c)$, and $(B \cap A) \cap (B \cap A^c) = \emptyset$. So, using Axiom (ii) for probability laws, $B \cap A = A$, and using Axiom (i) for probability laws,

$$\mathbf{P}(B) = \mathbf{P}(B \cap A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) \geq \mathbf{P}(A).$$

So, the first item is proven. We now prove the second item. Write $A = (A \setminus B) \cup (A \cap B)$ and note that $A \setminus B$ and $A \cap B$ are disjoint by Exercise 2.32. Similarly, write $B = (B \setminus A) \cup (B \cap A)$ and note that $(B \setminus A)$ and $(B \cap A)$ are disjoint. Finally, by Exercise 2.32, we can write $A \cup B$ as the union of three disjoint sets: $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$.

So, using Axiom (ii) for probability laws twice,

$$\mathbf{P}(A) + \mathbf{P}(B) = \mathbf{P}(A \setminus B) + \mathbf{P}(A \cap B) + \mathbf{P}(B \setminus A) + \mathbf{P}(A \cap B) = \mathbf{P}(A \cup B) + \mathbf{P}(A \cap B).$$

So, the second item is proven. The third and fourth items are left to the exercises. The final inequality follows from the third item and induction on n. □

Exercise 2.34. Let Ω be a sample space and let $\mathbf{P}$ be a probability law on Ω. Let $A, B, C \subseteq \Omega$. Show the following things:

- $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.
- $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

(**Although the book suggest otherwise, a Venn diagram alone is not a rigorous proof.** As in Exercise 2.17, a truth table allows us to rigorously reason about the information contained in a Venn diagram. Though, there are ways to do the problem while not directly using a truth table.)

2.3. **Conditional Probability.** Let's recall once again rolling a six-sided fair die. We can model this scenario with the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and with the uniform probability law $\mathbf{P}$ defined in Example 2.27. In particular, $\mathbf{P}(i) = 1/6$ for each $i \in \Omega$. That is, each of the six sides of the die has a $1/6$ chance of being rolled. However, suppose I roll the die out of your sight, and I just let you know that the die roll is even. Now, what is the probability of each roll? It is given that the die roll is even, so the probability that an odd number was rolled is 0. Since each even number was equally likely to be rolled, intuition suggests that each of the numbers $2, 4, 6$ has a probability of $1/3$ of being rolled. That is, the additional information that the roll was even has changed the probability of the events. Conditional probability allows us to compute the probabilities of events, given some previously unknown information.

Definition 2.35 (**Conditional Probability**)**.** Let A, B be subsets of some sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Assume that $\mathbf{P}(B) > 0$. We define the **conditional probability of A given B**, denoted by $\mathbf{P}(A|B)$, as

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

Example 2.36. Let $\Omega = \{1,2,3,4,5,6\}$ and let $\mathbf{P}$ be the uniform probability law on Ω. Let $B = \{2,4,6\}$. That is, B is the event that the die roll is even. If $i \in \Omega$ is odd, then $\{i\} \cap B = \emptyset$. So,

$$\mathbf{P}(\{i\}|B) = \frac{\mathbf{P}(\{i\} \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(\emptyset)}{1/2} = 0.$$

If $i \in \Omega$ is even, then $\{i\} \cap B = \{i\}$. So,

$$\mathbf{P}(\{i\}|B) = \frac{\mathbf{P}(\{i\} \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(\{i\})}{1/2} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Example 2.37. In Exercise 2.31, we considered $\Omega = \mathbb{R}^2$, and we defined the probability law

$$\mathbf{P}(A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dx dy, \qquad A \subseteq \Omega.$$

Let $B = \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 1\}$. Then B represents the circular dart board. So, if $A \subseteq \Omega$, then $P(A|B)$ is the probability that the dart lands in A, given than the dart has hit the dartboard. Using the definition of $\mathbf{P}$, we have

$$\mathbf{P}(A|B) = \frac{\iint_{A \cap B} e^{-(x^2+y^2)/2} dx dy}{\iint_B e^{-(x^2+y^2)/2} dx dy}.$$

In particular, $P(B|B) = 1$.

Proposition 2.38. *Let B be a fixed subset of some sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Assume that $\mathbf{P}(B) > 0$. Given any subset A in Ω, define $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$ as above. Then $\mathbf{P}(A|B)$ is itself a probability law on Ω.*

Proof. We first verify Axiom (i). Let $A \subseteq \Omega$. Since Axiom (i) holds for $\mathbf{P}$ by assumption, we have $\mathbf{P}(A \cap B) \geq 0$. Therefore, $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B) \geq 0$.

We now verify Axiom (iii). Note that $\mathbf{P}(\Omega|B) = \mathbf{P}(\Omega \cap B)/\mathbf{P}(B) = \mathbf{P}(B \cap B)/\mathbf{P}(B) = \mathbf{P}(B)/\mathbf{P}(B) = 1$.

We now verify Axiom (ii). Let $A, C \subseteq \Omega$ with $A \cap C = \emptyset$. Since A and C are disjoint, we know that $A \cap B$ and $C \cap B$ are disjoint. So, we can apply Axiom (ii) for $\mathbf{P}$ to the sets $A \cap B$ and $C \cap B$. So,

$$\begin{aligned}\mathbf{P}(A \cup C|B)\mathbf{P}(B) &= \mathbf{P}((A \cup C) \cap B) = \mathbf{P}((A \cap B) \cup (C \cap B)), \quad \text{by Proposition 2.16(ii)} \\ &= \mathbf{P}(A \cap B) + \mathbf{P}(C \cap B) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(C|B)\mathbf{P}(B).\end{aligned}$$

Dividing both sides by $\mathbf{P}(B)$ implies that Axiom (ii) holds for two sets. To verify that additivity holds for a countable number of sets, let $A_1, A_2, \ldots$ be subsets of Ω such that $A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \neq j$. Since $A_i \cap A_j = \emptyset$ whenever $i \neq j$, we have $(A_i \cap B) \cap (A_j \cap B) = \emptyset$. So, using Exercise 2.20, and Axiom (ii) for $\mathbf{P}$,

$$\begin{aligned}\mathbf{P}(B)\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k \,\middle|\, B\right) &= \mathbf{P}\left(\left(\bigcup_{k=1}^{\infty} A_k\right) \cap B\right) = \mathbf{P}\left(\bigcup_{k=1}^{\infty} (A_k \cap B)\right), \quad \text{by Exercise 2.20} \\ &= \sum_{k=1}^{\infty} \mathbf{P}(A_k \cap B) = \mathbf{P}(B) \sum_{k=1}^{\infty} \mathbf{P}(A_k|B)\end{aligned}$$

So, Axiom (ii) holds. In conclusion, $\mathbf{P}(A|B)$ is a probability law on Ω. $\square$

Remark 2.39. Proposition 2.38 implies that facts from Proposition 2.33 apply also to conditional probabilities. For example, using the notation of Proposition 2.38, we have $\mathbf{P}(A \cup C|B) \leq \mathbf{P}(A|B) + \mathbf{P}(C|B)$.

Example 2.40 (Medical Testing). Suppose a test for a disease is 99% accurate. That is, if you have the disease, the test will be positive with 99% probability. And if you do not have the disease, the test will be negative with 99% probability. Suppose also the disease is fairly rare, so that roughly 1 in $10,000$ people have the disease. If you test positive for the disease, with what probability do you actually have the disease?

The answer is unfortunately around 1/100. To see this, let's consider the probabilities. Let B be the event that you test positive for the disease. Let A be the event that you actually have the disease. We want to compute $\mathbf{P}(A|B)$. We have

$$\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B) = (\mathbf{P}(A)/\mathbf{P}(B))\mathbf{P}(A \cap B)/\mathbf{P}(A) = (\mathbf{P}(A)/\mathbf{P}(B))\mathbf{P}(B|A).$$

We are given that $\mathbf{P}(A) = 10^{-4}$, $P(B|A) = .99$ and $\mathbf{P}(B|A^c) = .01$. To compute $\mathbf{P}(B)$, we write $B = (B \cap A) \cup (B \cap A^c)$, so that

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(B \cap A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)\mathbf{P}(A^c) \\ &= .99(10^{-4}) + .01(1 - \mathbf{P}(A)) = .99(10^{-4}) + .01(1 - 10^{-4}) \approx 10^{-2}.\end{aligned}$$

In conclusion,

$$\mathbf{P}(A|B) = \frac{10^{-4}}{\mathbf{P}(B)}(.99) \approx 10^{-4}10^{2} = 10^{-2}.$$

So, even though the test is fairly accurate from a certain perspective, a positive test result does not say very much.

Many people find this result counterintuitive, though the following reasoning can help to explain the result. Suppose we have a population of $10,000$ people. Then roughly 1 person in the population has the disease. Suppose everyone is given the test. Since $9,999$ people are healthy and the test is 99% accurate, around 100 healthy people will test positive for the disease. Meanwhile, the 1 sick person will most likely test positive for the disease. So, out of around 101 people testing positive for the disease, only 1 of them actually has the disease. So, $\mathbf{P}(A|B)$ is roughly $1/101 \approx 10^{-2}$.

Some of the calculations from the previous problem will be formalized further below.

Example 2.41. Sometimes, conditioning on an event does *not* change the probability of an event. Let $\Omega = \{0,1\} \times \{0,1\}$, and let $\mathbf{P}$ be the uniform discrete probability law on Ω. We can think of $\mathbf{P}$ as modelling the flipping of two distinct fair coins, so that 0 denotes tails, and 1 denotes heads. Let $A = \{(0,1),(1,1)\}$, and let $B = \{(1,0),(1,1)\}$. Then B is the event that the first coin lands heads, and A is the event that the second coin lands heads. Note that $\mathbf{P}(A) = \mathbf{P}(B) = 1/2$. We compute $\mathbf{P}(A|B)$ as follows.

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(1,1)}{\mathbf{P}(B)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

That is, $\mathbf{P}(A|B) = \mathbf{P}(A)$. That is, given that the first coin lands heads, this does not affect the probability of the second coin landing heads. It is a common misconception (especially among gamblers) that, if a fair coin is flipped resulting in a tails, then a head "should" consequently occur on the next coin flip. However, this intuition is incorrect. We will return to examples of repeated experiments in our discussion of independence.

Proposition 2.42 (**Multiplication Rule**). *Let n be a positive integer. Let $A_1, \ldots, A_n$ be sets in some sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that $\mathbf{P}(A_i) > 0$ for all $i \in \{1, \ldots, n\}$. Then*

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_2 \cap A_1)\cdots\mathbf{P}(A_n|\cap_{i=1}^{n-1} A_i).$$

Proof. Using the definition of conditional probability,

$$\begin{aligned}&\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_2 \cap A_1)\cdots\mathbf{P}(A_n|\cap_{i=1}^{n-1} A_i)\\ &\quad= \mathbf{P}(A_1)\frac{\mathbf{P}(A_1 \cap A_2)}{\mathbf{P}(A_1)}\frac{\mathbf{P}(A_1 \cap A_2 \cap A_3)}{\mathbf{P}(A_1 \cap A_2)}\cdots\frac{\mathbf{P}(\cap_{i=1}^{n} A_i)}{\mathbf{P}(\cap_{i=1}^{n-1} A_i)} = \mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right).\end{aligned}$$

□

Exercise 2.43. Two fair coins are flipped. It is given that at least one of the coins is heads. What is the probability that the first coin is heads? (A flipped fair coin has either heads with probability 1/2, or tails with probability 1/2. In the real world, a coin has a small probability of landing on its side, but we are ignoring this possibility!)

Exercise 2.44 (**The Monty Hall Problem**). This Exercise demonstrates the sometimes counterintuitive nature of conditional probabilities.

You are a contestant on a game show. There are three doors labelled $1, 2$ and 3. You and the host are aware that one door contains a prize, and the two other doors have no prize. The host knows where the prize is, but you do not. Each door is equally likely to contain a prize, i.e. each door has a $1/3$ chance of containing the prize. In the first step of the game, you can select one of the three doors. Suppose the selected door is $i \in \{1, 2, 3\}$. Given your selection, the host now reveals one of the two remaining doors, demonstrating that this door contains no prize. The game now concludes with a choice. You can either keep your current door i, or you can switch to the other unopened door. You receive whatever is behind your selected door. The question is: should you switch or not?

If you do not switch your door choice, show that your probability of getting the prize at the end of the game is $1/3$. If you *do* switch your door choice, show that your probability of getting the prize is $2/3$. In conclusion, in this game, you should always switch your choice of doors.

2.4. **Total Probability Theorem.** In Example 2.40, we used the identity

$$\mathbf{P}(B) = \mathbf{P}(B \cap A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)\mathbf{P}(A^c).$$

This identity helped us to compute the probability of the event B. The Total Probability Theorem is a generalization of this fact.

Theorem 2.45 (**Total Probability Theorem**). *Let $A_1, \ldots, A_n$ be disjoint events in a sample space Ω. That is, $A_i \cap A_j = \emptyset$ whenever $i, j \in \{1, \ldots, n\}$ satisfy $i \neq j$. Assume also that $\cup_{i=1}^{n} A_i = \Omega$. Let $\mathbf{P}$ be a probability law on Ω. Then, for any event $B \subseteq \Omega$, we have*

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B \cap A_i) = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{P}(B|A_i).$$

Proof. We claim that $B = \cup_{i=1}^n (B \cap A_i)$, and the sets $B \cap A_1, \ldots, B \cap A_n$ are disjoint. Given this Claim, the result then follows from Remark 2.24. So, let's prove the claim.

We first show $B \subseteq \cup_{i=1}^n (B \cap A_i)$. Let $x \in B$. Then $x \in \Omega$ since $B \subseteq \Omega$. Since $\cup_{i=1}^n A_i = \Omega$, there exists $k \in \{1, \ldots, n\}$ such that $x \in A_k$. Since $x \in B$ as well, we conclude that $x \in B \cap A_k$. Therefore, $x \in \cup_{i=1}^n (B \cap A_i)$. We conclude that $B \subseteq \cup_{i=1}^n (B \cap A_i)$. We now show that $B \supseteq \cup_{i=1}^n (B \cap A_i)$. Since $B \cap A_i \subseteq B$ for all $i \in \{1, \ldots, n\}$, we have $\cup_{i=1}^n (B \cap A_i) \subseteq B$, as desired. In conclusion, $B = \cup_{i=1}^n (B \cap A_i)$.

It remains to show that the sets $B \cap A_1, \ldots, B \cap A_n$ are disjoint. Let $i, j \in \{1, \ldots, n\}$ with $i \neq j$. We need to show that $(B \cap A_i) \cap (B \cap A_j) = \emptyset$. By assumption, $A_i \cap A_j = \emptyset$. Therefore, $(B \cap A_i) \cap (B \cap A_j) = (A_i \cap A_j) \cap B = \emptyset$, as desired. □

Theorem 2.45 allows us to compute a potentially complicated probability by breaking it up into smaller sub-quantities. That is, Theorem 2.45 is most useful when $\mathbf{P}(B)$ is difficult to compute directly, and when we can find disjoint sets $A_1, \ldots, A_n$ with $\cup_{i=1}^n A_i = \Omega$ such that $\mathbf{P}(A_i \cap B)$ is easier to compute, for each $i \in \{1, \ldots, n\}$.

Example 2.46. Let's return to our example of rolling fair six-sided dice in Example 2.29, since we have already implicitly used Theorem 2.45 in our calculation there. Let $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$. Let $\mathbf{P}$ be the uniform probability law on Ω. Then $\mathbf{P}$ models the roll of two distinct, fair six-sided dice. Let B be the event that the sum of the dice is 5. We compute $\mathbf{P}(B)$ by conditioning on the identity of the first roll. For each $i \in \{1, 2, 3, 4, 5, 6\}$, let A_i be the event that the first roll is i. Then $\cup_{i=1}^6 A_i = \Omega$, since if $(j, k) \in \Omega$, then $(j, k) \in A_j$, so that $\Omega \subseteq \cup_{i=1}^6 A_i$, and also $\Omega \subseteq \cup_{i=1}^6 A_i$. Also, $A_i \cap A_j = \emptyset$ for all $i, j \in \{1, 2, 3, 4, 5, 6\}$ with $i \neq j$, since the events A_i and A_j exclude each other.

If $i < 5$, then $B \cap A_i = (i, 5 - i)$. (The first die is i and the sum of the dice is 5, so there is only one roll that the second die could have, namely $5 - i$). If $i > 5$, then $B \cap A_i = \emptyset$. So, using Theorem 2.45 and the definition of $\mathbf{P}$,

$$\mathbf{P}(B) = \sum_{i=1}^{6} \mathbf{P}(B \cap A_i) = \sum_{i=1}^{4} \mathbf{P}(B \cap A_i) = 4/36 = (1/36) \min(5 - 1, 13 - 5) = 1/9.$$

Exercise 2.47. Suppose you roll three distinct fair, four-sided dice. What is the probability that the sum of the dice is 7?

Exercise 2.48. Two people take turns throwing darts at a board. Person A goes first, and each of her throws has a probability of $1/4$ of hitting the bullseye. Person B goes next, and each of her throws has a probability of $1/3$ of hitting the bullseye. Then person A goes, and so on. With what probability will Person A hit the bullseye before Person B does?

Exercise 2.49. Suppose you roll two distinct fair six-sided dice. Suppose you roll these two dice again. What is the probability that both rolls have the same sum?

2.5. **Bayes' Rule.** In Example 2.40, we used the identity

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A)}{\mathbf{P}(B)} \mathbf{P}(B|A) = \frac{\mathbf{P}(A)\mathbf{P}(B|A)}{\mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)\mathbf{P}(A^c)}.$$

This identity helped us to compute $\mathbf{P}(A|B)$, since we were then able to compute the right side of the equality. Bayes' rule is a generalization of this fact. Bayes' rule allows us to reverse the order of the conditioning. That is, we want to compute a probability conditioned on B, but we instead compute probabilities conditioned on other events.

Theorem 2.50 (**Bayes' Rule**). *Let $A_1, \dots, A_n$ be disjoint events in a sample space Ω. That is, $A_i \cap A_j = \emptyset$ whenever $i, j \in \{1, \dots, n\}$ satisfy $i \neq j$. Assume also that $\cup_{i=1}^n A_i = \Omega$. Let $\mathbf{P}$ be a probability law on Ω. Then, for any event $B \subseteq \Omega$ with $\mathbf{P}(B) > 0$, and for any $j \in \{1, \dots, n\}$, we have*

$$\mathbf{P}(A_j|B) = \frac{\mathbf{P}(A_j)\mathbf{P}(B|A_j)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_j)\mathbf{P}(B|A_j)}{\sum_{i=1}^n \mathbf{P}(A_i)\mathbf{P}(B|A_i)}.$$

Proof. If $\mathbf{P}(A_j) = 0$, then both sides are zero. So, we may assume that $\mathbf{P}(A_j) > 0$. As in Example 2.40, we use Definition 2.35 to compute

$$\mathbf{P}(A_j|B) = \frac{\mathbf{P}(A_j \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_j)}{\mathbf{P}(A_j)} \frac{\mathbf{P}(A_j \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_j)}{\mathbf{P}(B)} \mathbf{P}(B|A_j).$$

That is, we proved the first equality. For the second equality, we apply Theorem 2.45 to the denominator term $\mathbf{P}(B)$. □

Exercise 2.51. Around 5% of men are colorblind, and around .25% of women are colorblind. Given that someone is colorblind, what is the probability that they are a man?

Exercise 2.52. Two people are flipping fair coins. Let n be a positive integer. Person I flips $n+1$ coins. Person II flips n coins. Show that the following event has probability 1/2: Person I has more heads than Person II.

2.6. **Recursions.** Some probabilities can be computed using self-referential, or recursive, equalities.

Example 2.53 (**Gambler's Ruin**). Let $0 < p < 1$. Suppose you are playing a game of chance. For each round of the game, with probability p you win \$1 and with probability $1-p$ you lose \$1. Suppose you start with \$50 and you decide to quit playing when you reach either \$0 or \$100. With what probability will you end up with \$100?

It is helpful to solve a more general problem, where 50 is replaced by any integer between 0 and 100. For each $i \in \{0, 1, 2, \dots, 100\}$, let B_i denote the event that you end up with \$100 if you started with \$$i$, and let $p_i = \mathbf{P}(B_i)$. So, we can at least determine that $p_0 = 0$ and $p_{100} = 1$. By conditioning on the result of the first round of the game, we can find a relation between the different p_i values.

Let A_1 be the event that the you win \$1 in the first round of the game, and let A_2 be the event that you lose \$1 in the first round of the game. Then $A_1 \cap A_2 = \emptyset$, and $A_1 \cup A_2 = \Omega$. Given that A_1 occurs, the probability of ending up with \$100 is the same as starting round one with one extra dollar. That is, $\mathbf{P}(B_i|A_1) = \mathbf{P}(B_{i+1})$. Similarly, $\mathbf{P}(B_i|A_2) = \mathbf{P}(B_{i-1})$. So, using Theorem 2.45,

$$\begin{aligned} p_i = \mathbf{P}(B_i) &= \mathbf{P}(A_1)\mathbf{P}(B_i|A_1) + \mathbf{P}(A_2)\mathbf{P}(B_i|A_2) \\ &= p\mathbf{P}(B_{i+1}) + (1-p)\mathbf{P}(B_{i-1}) = p \cdot p_{i+1} + (1-p)p_{i-1}. \end{aligned}$$

That is,

$$p_i = p \cdot p_{i+1} + (1-p) \cdot p_{i-1}, \qquad \forall\, i \in \{1, 2, \dots, 99\}.$$

That is, $p_{i-1} = (1-p)^{-1}(p_i - p \cdot p_{i+1})$. Or, written in matrix form

$$\begin{pmatrix} p_i \\ p_{i-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -p(1-p)^{-1} & (1-p)^{-1} \end{pmatrix} \begin{pmatrix} p_{i+1} \\ p_i \end{pmatrix}.$$

Iteratively applying this equation for any positive integer k, we have

$$\begin{pmatrix} p_i \\ p_{i-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -p(1-p)^{-1} & (1-p)^{-1} \end{pmatrix}^k \begin{pmatrix} p_{i+k} \\ p_{i+k-1} \end{pmatrix}.$$

Note that the eigenvalues λ of the matrix $\begin{pmatrix} 0 & 1 \\ -p(1-p)^{-1} & (1-p)^{-1} \end{pmatrix}$ satisfy $(-\lambda)((1-p)^{-1}-\lambda)+p(1-p)^{-1}=0$, so that $\lambda^2-(1-p)^{-1}\lambda+p(1-p)^{-1}=0$, so $\lambda^2(1-p)-\lambda+p=0$, so

$$\lambda = \frac{1 \pm \sqrt{1-4p(1-p)}}{2(1-p)} = \frac{1 \pm \sqrt{(2p-1)^2}}{2(1-p)} = \frac{1 \pm |2p-1|}{2(1-p)}.$$

That is, $\lambda = 1$ or $\lambda = p(1-p)^{-1}$. So, we see that the eigenvectors of the matrix are

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 1 \\ p(1-p)^{-1} \end{pmatrix}.$$

These vectors form a basis of $\mathbb{R}^2$ as long as $p \neq 1/2$. So, if $p \neq 1/2$, any column vector $x \in \mathbb{R}^2$ can be written as a linear combination of the form $x = \alpha v_1 + \beta v_2$. In particular, if $x = \begin{pmatrix} p_{100} \\ p_{99} \end{pmatrix}$, and if $x = \alpha v_1 + \beta v_2$, with $\alpha, \beta \in \mathbb{R}$, then

$$\begin{aligned} \begin{pmatrix} p_i \\ p_{i-1} \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ -p(1-p)^{-1} & (1-p)^{-1} \end{pmatrix}^{100-i} \begin{pmatrix} p_{i+(100-i)} \\ p_{i+(100-i-1)} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ -p(1-p)^{-1} & (1-p)^{-1} \end{pmatrix}^{100-i} \begin{pmatrix} p_{100} \\ p_{99} \end{pmatrix} = \alpha v_1 + \beta (p(1-p)^{-1})^{100-i} v_2. \end{aligned}$$

Since $p_0 = 0$, we have (using $i = 1$ above)

$$\begin{pmatrix} p_1 \\ 0 \end{pmatrix} = \alpha v_1 + \beta (p(1-p)^{-1})^{99} v_2.$$

So, $\alpha + \beta(p(1-p)^{-1})^{100} = 0$. Also, since $p_{100} = 1$ and $x = \alpha v_1 + \beta v_2$, we have $1 = \alpha + \beta$. Solving for α, β, we get

$$\beta \left(\frac{p}{1-p} \right)^{100} - \beta = -1.$$

That is, $\beta = 1/(1-(p(1-p)^{-1})^{100})$. And $\alpha = 1-\beta$. In conclusion, if $p \neq 1/2$,

$$\begin{aligned} p_i &= \alpha + \beta(p(1-p)^{-1})^{100-i} = 1 + [1/(1-(p(1-p)^{-1})^{100})][-1 + (p(1-p)^{-1})^{100-i}] \\ &= 1 + \frac{\left(\frac{1-p}{p}\right)^{100}}{\left(\frac{1-p}{p}\right)^{100} - 1}[-1 + (p(1-p)^{-1})^{100-i}] = 1 + \frac{-\left(\frac{1-p}{p}\right)^{100} + \left(\frac{1-p}{p}\right)^{i}}{\left(\frac{1-p}{p}\right)^{100} - 1} = \frac{\left(\frac{1-p}{p}\right)^{i} - 1}{\left(\frac{1-p}{p}\right)^{100} - 1}. \end{aligned}$$

To extend this equality to $p = 1/2$, we set $a = (1-p)/p$, let j be a positive integer, and use the identity $a^j - 1 = (a-1)(a^{j-1} + a^{j-2} + \cdots + a + 1)$ to get

$$p_i = \frac{\left(\frac{1-p}{p}\right)^{i} - 1}{\left(\frac{1-p}{p}\right)^{100} - 1} = \frac{\sum_{k=0}^{i-1} \left(\frac{1-p}{p}\right)^{k}}{\sum_{k=0}^{99} \left(\frac{1-p}{p}\right)^{k}}.$$

Letting $p \to 1/2$ shows that when $p = 1/2$, we get $p_i = i/100$.

2.7. **Independence of Sets.** Up until now, much of the material we have discussed could equally well have been understood through basic calculus or combinatorics. However, the first new concept that is specific to probability theory itself is the notion of independence. If I roll two fair dice, then the outcome of the first die roll does not depend at all on the outcome of the second die roll. That is, the die rolls are independent of each other. We formalize this notion in the following definition.

Definition 2.54 (**Independent Sets**)**.** Let A, B be subsets of a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that A and B are **independent** if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

Remark 2.55. If $\mathbf{P}(B) > 0$, and if A, B are independent, then $\mathbf{P}(A|B) = \mathbf{P}(A)$. That is, knowing the event B does not affect the probability of A occurring.

Example 2.56. Let's return once again to our example of rolling two fair dice. Let $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$, and let $\mathbf{P}$ be the uniform probability law on Ω. Let $i, j \in \{1,2,3,4,5,6\}$. Let A_i be the event that the first die roll is i, and let B_j be the event that the second die roll is j. Then, for every $i, j \in \{1,2,3,4,5,6\}$, the events A_i and B_j are independent. That is, the one die roll does not affect the other die roll at all. To see this, note that $\mathbf{P}(A_i) = \mathbf{P}(B_j) = 1/6$, while $A_i \cap B_j$ is the single element $(i, j) \in \Omega$. That is, $\mathbf{P}(A_i \cap B_j) = 1/36$. Therefore,

$$\mathbf{P}(A_i \cap B_j) = 1/36 = (1/6)^2 = \mathbf{P}(A_i)\mathbf{P}(B_j).$$

Remark 2.57. In a probabilistic model, when two actions do not really affect each other (such as rolling two fair die), then we can anticipate independence in the model. However, there are many times when independence is not a valid assumption, and it is important to note when this is true. For example, suppose we let A be the event that one voter votes for candidate Alice, and let A' be the event that another voter votes for candidate Alice. If two voters are friends, or they watch the same news media, etc., then the events A, A' will probably not be independent. For other examples, consider the recession in the stock market in August of 2008. Many people believe that the following scenario caused the crash: several financial models all assumed that each financial entity was acting independently. However, in reality, many financial entities were using similar or identical models to decide which stocks to buy and sell. So, the entities were not acting independently at all! Since the models were wrong, they automatically made bad decisions, causing money to evaporate very quickly.

Definition 2.58 (**Independent Sets**)**.** Let n be a positive integer. Let $A_1, \ldots, A_n$ be subsets of a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that $A_1, \ldots, A_n$ are **independent** if, for any subset S of $\{1, \ldots, n\}$, we have

$$\mathbf{P}\left(\cap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i).$$

Remark 2.59. Note that the above definition is much stronger than simply requiring that $\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j)$ for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$, since the latter condition corresponds only to subsets S of $\{1, \ldots, n\}$ of size at most 2. In fact, the condition $\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j)$ for all $i \neq j$ does *not* imply that all of the sets are independent, as we now show by counterexample.

Example 2.60. Let $\Omega = \{H, T\} \times \{H, T\}$. Then Ω is a sample space representing two separate coin flips (H stands for heads, and T stands for tails). Let $\mathbf{P}$ denote the uniform probability law on Ω. Let A_1 be the event that the first coin toss is H (heads). Let A_2 be the event that the second coin toss is H (heads). Let A_3 be the event that both coin tosses are different. We will show that the events A_1, A_2, A_3 are pairwise independent, but they are not independent. That is, $\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j)$ for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$, but these three sets are not independent.

Note that $\mathbf{P}(A_1) = \mathbf{P}(A_2) = 1/2$ and $A_1 \cap A_2 = (H, T)$, so

$$\mathbf{P}(A_1 \cap A_2) = 1/4 = (1/2)^2 = \mathbf{P}(A_1)\mathbf{P}(A_2).$$

Also, $\mathbf{P}(A_3) = 1/2$, $A_1 \cap A_3 = (H, T)$ and $A_2 \cap A_3 = (T, H)$, so

$$\mathbf{P}(A_1 \cap A_3) = 1/4 = (1/2)^2 = \mathbf{P}(A_1)\mathbf{P}(A_3), \qquad \mathbf{P}(A_2 \cap A_3) = 1/4 = (1/2)^2 = \mathbf{P}(A_2)\mathbf{P}(A_3).$$

In conclusion, each pair of the events A_1, A_2, A_3 are independent. That is, the definition of independence holds for any subset $S \subseteq \{1, 2, 3\}$ of size two. However, the definition of independence *fails* when $S = \{1, 2, 3\}$. Indeed, $A_1 \cap A_2 \cap A_3 = \emptyset$, so that

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(\emptyset) = 0 \neq 1/8 = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3).$$

So, the events A_1, A_2, A_3 are *not* independent.

Proposition 2.61. *Let A, B be subsets of a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that A and B are independent. Then A and B^c are independent.*

Proof. Writing $A = (A \cap B) \cup (A \cap B^c)$ where the union is disjoint, we have

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A) - \mathbf{P}(A \cap B) = \mathbf{P}(A) - \mathbf{P}(A)\mathbf{P}(B) = \mathbf{P}(A)(1 - \mathbf{P}(B)) = \mathbf{P}(A)\mathbf{P}(B^c).$$

□

The following two Exercises show that independence of sets can sometimes have a geometric interpretation.

Exercise 2.62. Let $\Omega = [0, 1] \times [0, 1]$ so that $\Omega \subseteq \mathbb{R}^2$. Define a probability law $\mathbf{P}$ so that, for any set $A \subseteq \Omega$, $\mathbf{P}(A)$ is defined to be the area of A. Let $0 \leq a_1 \leq a_2 \leq 1$ and let $0 \leq b_1 \leq b_2 \leq 1$. Consider the rectangles $A = \{(x, y) \in \Omega \colon a_1 \leq x \leq a_2\}$, $B = \{(x, y) \in \Omega \colon b_1 \leq y \leq b_2\}$. Show that the rectangles A, B are independent.

Exercise 2.63. Let $\Omega = \mathbb{R}^2$ so that $\Omega \subseteq \mathbb{R}^2$. Define a probability law $\mathbf{P}$ so that, for any set $A \subseteq \Omega$,

$$\mathbf{P}(A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dx dy.$$

In Exercise 2.31, we verified that $\mathbf{P}(\Omega) = 1$. Let $0 \leq a_1 \leq a_2 \leq 1$ and let $0 \leq b_1 \leq b_2 \leq 1$. Consider the infinite rectangles $A = \{(x, y) \in \Omega \colon a_1 \leq x \leq a_2\}$, $B = \{(x, y) \in \Omega \colon b_1 \leq y \leq b_2\}$. Show that the rectangles A, B are independent.

Example 2.64 (Bernoulli Trials). Let n be a positive integer. Let $\Omega = \{H, T\}^n$. Then Ω is a sample space representing n separate coin flips (H stands for heads, and T stands for tails). Let $0 < p < 1$. Let $\mathbf{P}$ be the probability law such that each coin toss occurs independently, and such that each coin has probability p of heads (H), and probability $1 - p$ of tails (T). That is, we are independently flipping n biased coins.

Let $1 \leq k \leq n$. Suppose the first k coins have landed as heads, and the rest of the coins are tails. By the definition of $\mathbf{P}$, this event occurs with probability $p^k(1-p)^{n-k}$. We now ask: What is the probability that k of the coins are heads, and the remaining $n-k$ coins are tails? In order to answer this question, we need to compute $C_{n,k}$, the number of unordered lists of k copies of H, and $n-k$ copies of T. Equivalently, $C_{n,k}$ is the number of ways to place n coins on a table all showing tails, and then turn over k distinct coins to reveal exactly k heads. To compute the latter number, note that we can first turn over one of the n coins, and then we can turn over any of the remaining $n-1$ coins showing tails, and then we can turn over any of the remaining $n-2$ coins showing tails, and so on. So, there are $n(n-1)(n-2)\cdots(n-k+1)$ sequences of coin turns which can be made (while keeping track of their ordering). To make the same count of coin flips without keeping track of the ordering, we just divide by the number of orderings of the k heads coins, which is $k(k-1)\cdots(2)(1)$. In conclusion,

$$C_{n,k} = \binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots(2)(1)} = \frac{n!}{(n-k)!k!}.$$

Back to our original question, the probability that we have k heads and $n-k$ tails among n coin flips is

$$C_{n,k} \cdot p^k(1-p)^{n-k} = \binom{n}{k} p^k(1-p)^{n-k} = \frac{n!}{(n-k)!k!} p^k(1-p)^{n-k}.$$

Theorem 2.65 (**Binomial Theorem**). *Let $0 < p < 1$. Then*

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

More generally, for any real numbers x, y, we have

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

Proof. We use the notation of Example 2.64. Let $0 < p < 1$. For any $0 \leq k \leq n$, let A_k be the event that there are exactly k heads that resulted from flipping n coins. Then $A_i \cap A_j = \emptyset$ for all $i \neq j$ where $i, j \in \{0, \ldots, n\}$. Also, $\cup_{k=0}^{n} A_k = \Omega$. From Example 2.64, $\mathbf{P}(A_k) = \binom{n}{k} p^k(1-p)^k$. So, using Remark 2.24,

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\cup_{k=0}^{n} A_k) = \sum_{k=0}^{n} \mathbf{P}(A_k) = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k}. \qquad (*)$$

Now, the right side is a polynomial in p, which is equal to 1 for all $0 < p < 1$. Therefore, the equality $(*)$ holds for all real p. (A polynomial which is equal to 1 on $[0,1]$ is also equal to 1 on the whole real line.) Assume temporarily that $x + y \neq 0$. Define $p = x/(x+y)$. Then $x = p(x+y)$, $y = (1-p)(x+y)$ and $1 - p = y/(x+y)$. Using $(*)$, we have

$$1 = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{x}{x+y}\right)^k \left(\frac{y}{x+y}\right)^{n-k} = (x+y)^{-n} \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

That is, our desired equality holds whenever $x + y \neq 0$. Finally, the case $x + y = 0$ follows by letting $t > 0$ be a real parameter, using $\sum_{k=0}^{n} \binom{n}{k} x^k (y+t)^{n-k} = (x+y+t)^n$, and letting $t \to 0$, noting that both sides of the equality are continuous in t. □

Exercise 2.66. Let Ω be a sample space and let $\mathbf{P}$ be a probability law on Ω. Let $A, B \subseteq \Omega$. Assume that $A \subseteq B$. Is it possible that A is independent of B? Justify your answer.

2.8. **Counting Problems.** The following facts from counting are discussed in more detail in the Combinatorics class, Math 61.

Proposition 2.67 (**Counting Principles**). *Let n be a positive integer, and let k be an integer with $0 \leq k \leq n$. We define $n! = n \cdot (n-1) \cdot (n-2) \cdots (2) \cdot 1$.*

- *The number of permutations of the set $\{1, 2, \ldots, n\}$ is $n!$. That is, there are $n!$ ways to make an ordered list of the numbers $\{1, 2, \ldots, n\}$.*
- *The number of ways to make an ordered list of k elements of the set $\{1, 2, \ldots, n\}$ is $n!/(n-k)! = n(n-1)\cdots(n-k+1)$.*
- *The number of ways to make an unordered list of k elements of the set $\{1, 2, \ldots, n\}$ is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Equivalently, there are $\binom{n}{k}$ ways to partition the set $\{1, 2, \ldots, n\}$ into two parts such that one part contains exactly k elements.*
- *Let $n_1, \ldots, n_i$ be positive integers such that $n_1 + \cdots + n_i = n$. Then the number of ways to partition the set $\{1, \ldots, n\}$ into i sets, where the j^{th} group has n_j elements, for each $1 \leq j \leq i$, is*

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

Proof. We have essentially proven the first three facts in Example 2.64 □

Exercise 2.68 (**Inclusion-Exclusion Formula**). In the Properties for Probability laws, we showed that $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. The following equality is a generalization of this fact. Let Ω be a discrete sample space, and let $\mathbf{P}$ be a probability law on Ω. Prove the following. Let $A_1, \ldots, A_n \subseteq \Omega$. Then:

$$\mathbf{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \mathbf{P}(A_i \cap A_j \cap A_k) \\ \cdots + (-1)^{n+1} \mathbf{P}(A_1 \cap \cdots \cap A_n).$$

(Hint: begin with the identity $0 = (1-1)^m = \sum_{k=0}^{m} (-1)^k C_{m,k}$, which follows from the Binomial Theorem. That is, $1 = \sum_{k=1}^{m} (-1)^{k+1} C_{m,k}$. Now, let $x \in \Omega$ such that x is in exactly m of the sets $A_1, \ldots, A_n$. Compute the "number of times" that the element $x \in \Omega$ is counted for both sides of the Inclusion-Exclusion Formula.)

Exercise 2.69 (**Derangements**).

- Suppose you have a car with four tires, and the car mechanic removes all four tires. Suppose the mechanic now puts the tires back on randomly, so that all arrangements of the tires are equally likely. With what probability will no tire end up in its original position? (Hint: let A_i be the event that the i^{th} tire is in the correct position, where $i = 1, 2, 3, 4$. Then, use the Inclusion-Exclusion formula.)

- Let n be a positive integer. Suppose your car has n tires that are removed. Suppose the mechanic now puts the tires back on randomly, so that all arrangements of the tires are equally likely. With what probability will no tire end up in its original position?
- Compute the latter probability as $n \to \infty$.

3. Discrete Random Variables

So far we have discussed random events. Often it is also natural to describe random numbers. For example, the sum of two six-sided die is a random number. Or your score obtained by throwing a single dart at a standard dartboard is a random number. In probability, we call random numbers **random variables**.

Definition 3.1 (**Random Variable**)**.** Let Ω be a sample space. Let $\mathbf{P}$ be a probability law on Ω. A **random variable** X is a function $X\colon \Omega \to \mathbb{R}$. A **discrete random variable** is a random variable whose range is either finite or countably infinite.

Proposition 3.2 (**Properties of Random Variables**)**.**

- *If X and Y are random variables, then $X + Y$ is a random variable.*
- *If X is a random variable and if $f\colon \mathbb{R} \to \mathbb{R}$ is a function, then $f(X) = f \circ X$ is a random variable.*

A random variable is "just" a function. So, in some sense, from your preparation in calculus, you are already quite familiar with random variables. However, the new terminology of "random variable" carries a new perspective on functions as well. For example, in probability theory, we concern ourselves with the probability that the random variable takes various values.

Example 3.3. Let $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. Let $\mathbf{P}$ denote the uniform probability law on Ω. As usual, Ω and $\mathbf{P}$ denote the rolling of two distinct fair six-sided dice. We define random variables X, Y as follows. For any $(i, j) \in \Omega$, define $X(i, j) = i$, and define $Y(i, j) = j$. Then X and Y are random variables. Moreover, X is the roll of the first die, and Y is the roll of the second die. So, $X+Y$ is the sum of the rolls of the dice, and $X+Y$ is a random variable.

Example 3.4. Consider the following simplified version of a dartboard. Let $\Omega = \mathbb{R}^2$. For any set $A \subseteq \Omega$, define

$$\mathbf{P}(A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dx dy.$$

Let $(x, y) \in \Omega$. Define a random variable $X\colon \Omega \to \mathbb{R}$ so that

$$X(x, y) = \begin{cases} 1 & \text{, if } x^2 + y^2 \leq 1 \\ 0 & \text{, if } x^2 + y^2 > 1 \end{cases}.$$

That is, if you hit the dartboard $\{(x, y) \in \Omega\colon x^2 + y^2 \leq 1\}$, then $X = 1$. Otherwise, $X = 0$. So, X is a random variable which represents your score after throwing a random dart according to the probability law $\mathbf{P}$.

Example 3.5. Consider the following model of a more complicated dartboard. Let $\Omega = (0, 1)^2 \subseteq \mathbb{R}^2$. For any set $A \subseteq \Omega$, let $\mathbf{P}(A)$ denote the area of A. Let $(x, y) \in \Omega$. Define a random variable $X\colon \Omega \to \mathbb{R}$ so that $X(x, y)$ is the smallest integer j such that $x > 2^{-j}$

and $y > 2^{-j}$. For example, if $(x, y) = (1/3, 1/3)$, then $2^{-1} > x > 2^{-2}$ and $2^{-1} > y > 2^{-2}$, so $X(x, y) = 2$. Or if $(x, y) = (1/5, 1/3)$, then $2^{-2} > x > 2^{-3}$ and $2^{-1} > y > 2^{-2} > 2^{-3}$, so $X(x, y) = 3$. In this example, X is a random variable which represents your score after throwing a random dart according to the probability law $\mathbf{P}$. By the definition of X, if we would like to get a large score, we see that it is more beneficial to aim for the bottom left corner of the square, i.e. we want to get close to $(0, 0)$.

If we have a random variable X, one of the first tasks in probability is to compute various quantities for X to better understand X. For example, we could ask, "What value does X typically take?" (What is the mean value or average value of X?) "Typically, how far is X from its mean value?" (What is the variance of X?) We will start to answer these questions in Section 4. For now, we need to get through some preliminary concepts.

3.1. Probability Mass Function (PMF).

Definition 3.6 (**Probability Mass Function**). Let X be a random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x \in \mathbb{R}$. Consider the event $\{\omega \in \Omega\colon X(\omega) = x\}$. This event is often denoted as $\{X = x\}$. The **probability mass function** of X, denote $p_X\colon \mathbb{R} \to [0, 1]$ is defined by

$$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{X = x\}) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) = x\}), \qquad x \in \mathbb{R}.$$

Let $A \subseteq \mathbb{R}$. We denote $\{\omega \in \Omega\colon X(\omega) \in A\} = \{X \in A\}$.

Example 3.7. Let $\Omega = \{H, T\}^2$ and let $\mathbf{P}$ be the uniform probability measure on Ω. Then Ω and $\mathbf{P}$ represent the outcome of flipping two distinct fair coins. Let X be the number of heads that are rolled. That is, $X(T, T) = 0$, $X(H, T) = 1$, $X(T, H) = 1$ and $X(H, H) = 2$. Therefore,

$$p_X(x) = \begin{cases} 1/4 & \text{, if } x = 0 \\ 1/2 & \text{, if } x = 1 \\ 1/4 & \text{, if } x = 2 \\ 0 & \text{, otherwise.} \end{cases}$$

Note that $\mathbf{P}(X > 0) = 1/2 + 1/4 = 3/4$. That is, with probability $3/4$, at least one head is rolled.

Proposition 3.8. *Let X be a discrete random variable on a sample space Ω. Then*

$$\sum_{x \in \mathbb{R}} p_X(x) = 1.$$

Proof. For each $x \in \mathbb{R}$, let B_x be the event that $X = x$. If $x \neq y$, then $B_x \cap B_y = \emptyset$. Also, $\cup_{x \in \mathbb{R}} B_x = \Omega$. So, using Axiom (ii) for probability laws in Definition 2.22,

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\cup_{x \in \mathbb{R}} B_x) = \sum_{x \in \mathbb{R}} \mathbf{P}(B_x) = \sum_{x \in \mathbb{R}} p_X(x).$$

□

We now give descriptions of some commonly encountered random variables.

Definition 3.9 (**Bernoulli Random Variable**)**.** Let $0 < p < 1$. A random variable X is called a **Bernoulli random variable with parameter** p if $X = 1$ with probability p, and $X = 0$ with probability $1 - p$. Put another way, $X = 1$ when a single flipped biased coin lands heads, and $X = 0$ when the coin lands tails. The PMF is given by

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

Remark 3.10. Note that we defined the random variable X without specifying any sample space Ω. This de-emphasis on the domain is one aspect of probability that we mentioned above. For example, we could choose $\Omega = \{0, 1\}$ and define $\mathbf{P}$ on Ω such that $\mathbf{P}(0) = 1 - p$ and $\mathbf{P}(1) = p$. Then define $X \colon \Omega \to \mathbb{R}$ so that $X(\omega) = \omega$ for all $\omega \in \Omega$. Then X is a Bernoulli random variable.

Alternatively, we could choose $\Omega = [0, 5]$, and define $\mathbf{P}$ on Ω such that $\mathbf{P}[a, b] = \frac{1}{5}(b - a)$ whenever $0 \leq a < b \leq 5$. Then, we could define $Y \colon \Omega \to \mathbb{R}$ by

$$Y(\omega) = \begin{cases} 1 & \text{, if } \omega < 5p \\ 0 & \text{, if } \omega \geq 5p. \end{cases}$$

Then Y is also a Bernoulli random variable. As we can see, the sample spaces of X and Y are very different.

Definition 3.11 (**Binomial Random Variable**)**.** Let $0 < p < 1$ and let n be a positive integer. A random variable X is called a **binomial random variable with parameters** n **and** p if X has the following PMF. If k is an integer with $0 \leq k \leq n$, then

$$p_X(k) = \mathbf{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

For any other x, we have $p_X(x) = 0$. In Example 2.64, we showed that this probability distribution arises from flipping n biased coins. In particular, X is the number of heads that arise when flipping n biased coins. In Theorem 2.65, we verified that

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

Definition 3.12 (**Geometric Random Variable**)**.** Let $0 < p < 1$. A random variable X is called a **geometric random variable with parameter** p if X has the following PMF. If k is a positive integer, then

$$p_X(k) = \mathbf{P}(X = k) = (1 - p)^{k-1} p.$$

For any other x, we have $p_X(x) = 0$. Note that X is the number of times that are needed to flip a biased coin in order to get a heads (if the coin has probability p of landing heads). Also, using the summation of geometric series, we verify

$$\begin{aligned} \sum_{k=1}^{\infty} p_X(k) &= \sum_{k=1}^{\infty} (1 - p)^{k-1} p = p \sum_{k=1}^{\infty} (1 - p)^{k-1} = p \lim_{n \to \infty} \sum_{k=1}^{n} (1 - p)^{k-1} \\ &= p \lim_{n \to \infty} \frac{1 - (1 - p)^{n+1}}{1 - (1 - p)} = \frac{p}{p} = 1. \end{aligned}$$

Definition 3.13 (**Poisson Random Variable**). Let $\lambda > 0$. A random variable X is called a **Poisson random variable with parameter** λ if X has the following PMF. If k is a nonnegative integer, then

$$p_X(k) = \mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

For any other x, we have $p_X(x) = 0$. Using the Taylor expansion for the exponential function, we verify

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

The Poisson random variable provides a nice approximation to the binomial distribution, as we now demonstrate.

Proposition 3.14 (**Poisson Approximation to the Binomial**). *Let $\lambda > 0$. For each positive integer n, let $0 < p_n < 1$, and let X_n be a binomial distributed random variable with parameters n and p_n. Assume that $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$. Then, for any nonnegative integer k, we have*

$$\lim_{n\to\infty} \mathbf{P}(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Lemma 3.15. *Let $\lambda > 0$. For each positive integer n, let $\lambda_n > 0$. Assume that $\lim_{n\to\infty} \lambda_n = \lambda$. Then*

$$\lim_{n\to\infty}\left(1 - \frac{\lambda_n}{n}\right)^n = e^{-\lambda}$$

Proof. Let log denote the natural logarithm. For any $x < 1$, define $f(x) = \log(1-x)$. From L'Hôpital's Rule,

$$\lim_{x\to 0}\frac{f(x)}{x} = \lim_{x\to 0} f'(x) = \lim_{x\to 0}\frac{-1}{1-x} = -1. \qquad (*)$$

So, using $\lim_{n\to\infty} \lambda_n/n = 0$ we can apply $(*)$ and then $\lim_{n\to\infty}\lambda_n = \lambda$, so

$$\begin{aligned}\lim_{n\to\infty}\left(1 - \frac{\lambda_n}{n}\right)^n &= \lim_{n\to\infty}\exp\left(\log\left(1 - \frac{\lambda_n}{n}\right)^n\right)\\ &= \exp\left(\lim_{n\to\infty}\frac{\log\left(1-\frac{\lambda_n}{n}\right)}{\lambda_n/n}\lambda_n\right) = \exp((-1)(\lambda)) = e^{-\lambda}.\end{aligned}$$

□

Proof of Proposition 3.14. For any positive integer n, let $\lambda_n = np_n$. Then $\lim_{n\to\infty}\lambda_n = \lambda$ and $\lim_{n\to\infty}\lambda_n/n = 0$. And if k is a nonnegative integer,

$$\begin{aligned}\mathbf{P}(X_n = k) &= \binom{n}{k}\left(\frac{\lambda_n}{n}\right)^k\left(1-\frac{\lambda_n}{n}\right)^{n-k}\\ &= \left(\prod_{i=1}^{k}\frac{n-i+1}{n}\right)\frac{\lambda_n^k}{k!}\left(1-\frac{\lambda_n}{n}\right)^n\left(1-\frac{\lambda_n}{n}\right)^{-k}\end{aligned}$$

So, using Lemma 3.15, $\lim_{n\to\infty}\mathbf{P}(X_n = k) = 1\cdot\frac{\lambda^k}{k!}e^{-\lambda}\cdot 1$. □

Remark 3.16. A Poisson random variable is often used as an approximation for counting the number of some random occurrences. For example, the Poisson distribution can model the number of typos per page in a book, the number of magnetic defects in a hard drive, the number of traffic accidents in a day, etc.

Exercise 3.17. The Wheel of Fortune involves the repeated spinning of a wheel with 72 possible stopping points. We assume that each time the wheel is spun, any stopping point is equally likely. Exactly one stopping point on the wheel rewards a contestant with $\$1,000,000$. Suppose the wheel is spun 24 times. Let X be the number of times that someone wins $\$1,000,000$. Using the Poisson Approximation the Binomial, estimate the following probabilities: $\mathbf{P}(X = 0)$, $\mathbf{P}(X = 1)$, $\mathbf{P}(X = 2)$. (Hint: consider the binomial distribution with $p = 1/72$.)

Remark 3.18. The Bernoulli, binomial, geometric and Poisson random variables are all examples of the following general construction of a random variable. Let $a_0, a_1, a_2, \ldots \geq 0$ such that $\sum_{i=0}^{\infty} a_i = 1$. Then define a random variable X such that $\mathbf{P}(X = i) = a_i$ for all nonnegative integers i.

There are many other random variables we will encounter in this class as well, but these will be enough for now.

3.2. **Functions of Random Variables.**

Proposition 3.19. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω, and let $f\colon \mathbb{R} \to \mathbb{R}$. Then $f(X)$ has PMF*

$$p_{f(X)}(y) = \sum_{x \in \mathbb{R}\colon\, f(x)=y} p_X(x), \qquad \forall\, y \in \mathbb{R}.$$

Proof. Let $x, y, z \in \mathbb{R}$. Let A_x be the event that $X = x$. If $z \neq x$, then $A_x \cap A_z = \emptyset$. Also, $\cup_{x \in \mathbb{R}} A_x = \Omega$. So, using Axiom (ii) of Definition 2.22,

$$p_{f(X)}(y) = \mathbf{P}(f(X) = y) = \mathbf{P}(\cup_{x \in \mathbb{R}} \{f(X) = y\} \cap A_x) = \sum_{x \in \mathbb{R}} \mathbf{P}(\{f(X) = y\} \cap A_x)$$

$$= \sum_{x \in \mathbb{R}\colon\, f(x)=y} \mathbf{P}(X = x) \;=\; \sum_{x \in \mathbb{R}\colon\, f(x)=y} p_X(x).$$

□

Exercise 3.20. Let $\Omega = \{-3, -2, -1, 0, 1, 2, 3\}$. Suppose $X(\omega) = \omega$ for all $\omega \in \Omega$. Let $f\colon \mathbb{R} \to \mathbb{R}$ so that $f(x) = x^2$ for any $x \in \mathbb{R}$. Compute the PMF of $f(X)$.

4. Expectation, Conditioning

Now that we understand random variables a bit more, we can finally start to answer some of the fundamental questions of probability, such as:

What is the average value of a random variable?

Put another way, what is the mean value of a random variable? Or, what value should we expect a particular random variable to have? Answering this question is of interest in many applications of probability. For example, if I can figure out a way to gain $\$1$ from a stock transaction with probability .51, while losing $\$1$ from a stock transaction with probability

.49, and if I keep performing this transaction many times, I should probably expect to gain money over time.

Example 4.1 (Playing Monopoly Forever). Suppose you are moving a game piece on a large monopoly board. At each turn, you roll a fair six-sided die, and you move the piece the number of spaces that is rolled. With what probability will you land on the space that is 40 spaces away from the starting point?

Let i be a positive integer. Let A_i be the event that you land on the location that is i spaces away from the starting location, after any number of rolls. Let $p_i = \mathbf{P}(A_i)$. By conditioning on the first roll, we can find a recurrence relation for the p_i. Let B_j be the event that the first die roll is j, where $j \in \{1,2,3,4,5,6\}$. If $i > 6$, then $\mathbf{P}(A_i|B_j) = \mathbf{P}(A_{i-j})$. Then by Theorem 2.45,

$$p_i = \mathbf{P}(A_i \cap \Omega) = \mathbf{P}(A_i \cap (\cup_{j=1}^{6} B_j)) = \sum_{j=1}^{6} \mathbf{P}(B_j)\mathbf{P}(A_i|B_j) = \frac{1}{6}\sum_{j=1}^{6} \mathbf{P}(A_{i-j}) = \frac{1}{6}\sum_{j=1}^{6} p_{i-j}. \quad (*)$$

We could theoretically solve this recursion as in Example 2.53, but for simplicity, we will instead just compute $p_1, \ldots, p_6$ directly, and then compute p_i for $i > 6$ using the recursion $(*)$.

The only way to land on the first space is to roll a 1 on the first roll, so $p_1 = 1/6$. We can land on the second space by rolling a two on the first roll, or by rolling two consecutive 1's. So, $p_2 = 1/6 + (1/6)^2$. We can land on the third space by rolling: a 3; three 1's; or one 2 and one 1. So, $p_3 = 1/6 + (1/6)^3 + 2(1/6)^2$. Similarly, $p_4 = 1/6 + (1/6)^4 + 3(1/6)^3 + 2(1/6)^2 + (1/6)^2$, and so on. Here is a table showing the values of $p_1, p_2, \ldots, p_{40}$, where the first eight values descend in the first column, then the nest eight values descend in the second column, etc.

0.166666666667	0.280368945441	0.286701924733	0.285599870541	0.285721682947
0.194444444444	0.289288461040	0.285586725149	0.285747713887	0.285713826853
0.226851851852	0.293393122242	0.284712810463	0.285768819510	0.285710193751
0.264660493827	0.290830213260	0.285621080152	0.285735625468	0.285711733841
0.308770576132	0.279263192334	0.285967983759	0.285700953208	0.285715051280
0.360232338820	0.283539658507	0.285943659029	0.285691829207	0.285716315054
0.253604395290	0.286113932137	0.285755697214	0.285707468637	0.285714800621
0.268094016728	0.287071429920	0.285597992628	0.285725401653	0.285713653567

It looks like the sequences of numbers $p_1, p_2, \ldots$ is converging to something. If we continue this computation we get $p_{100} = 0.285714285714$. That is, $p_{100} \approx 2/7$. Why is this so?

For each $i \geq 1$, let X_i denote the result of the i^{th} die roll. Then $X_1 + \cdots + X_n$ is the number of spaces that is moved after n rolls of the dice. How many spaces can we expect to move after a single die roll? If $i \geq 1$, then X_i can be any of the numbers $\{1,2,3,4,5,6\}$ with equal probability. If n is very large, and if we interpret probabilities as frequencies, then around $1/6$ of the indices $i \in \{1,\ldots,n\}$ satisfy $X_i = 1$, around $1/6$ of the indices $i \in \{1,\ldots,n\}$ satisfy $X_i = 2$, and so on. That is, when n is large,

$$X_1 + \cdots + X_n \approx \frac{n}{6}(1) + \frac{n}{6}(2) + \frac{n}{6}(3) + \frac{n}{6}(4) + \frac{n}{6}(5) + \frac{n}{6}(6) = \frac{n}{6}(1+2+3+4+5+6) = n\frac{21}{6} = n\frac{7}{2}.$$

Written another way,

$$\frac{X_1 + \cdots + X_n}{n} \approx \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = \frac{7}{2}.$$

That is, on average, each die roll will move forward around 7/2 spaces. Put another way, after two rolls, we will have visited two spaces while moving forward around seven spaces. That is, we will have visited two spaces and skipped five in between. So, the probability of landing on any particular space is $2/7 = 1/(7/2)$.

In the above example, we reasoned that, on average, we can expect the roll of a single fair die to be around 7/2. This fact is formalized by defining the expected value of a random variable.

4.1. **Expectation, Variance.**

Definition 4.2 (**Expected Value**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. Assume that $X\colon \Omega \to [0,\infty)$. We define the **expected value** of X, denoted $\mathbf{E}(X)$, by

$$\mathbf{E}(X) = \sum_{x\in\mathbb{R}} x p_X(x).$$

For a discrete random variable with $X\colon \Omega \to \mathbb{R}$, if $\mathbf{E}\,|X| < \infty$, we then define $\mathbf{E}(X) = \sum_{x\in\mathbb{R}} x p_X(x)$ as above. The expected value of X is also referred to as the **mean** of X, or the **first moment** of X. More generally, if n is a positive integer, we define the n^{th} **moment** of X to be $\mathbf{E}(X^n)$.

Example 4.3. If X takes the values $\{1,2,3,4,5,6\}$ each with probability 1/6, then we have already verified in Example 4.1 that

$$\mathbf{E}(X) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = \frac{21}{6} = \frac{7}{2}.$$

That is, on average, the result of the roll of one fair six-sided die will be around 7/2. We can also compute

$$\mathbf{E}(X^2) = \frac{1}{6}(1^2) + \frac{1}{6}(2^2) + \frac{1}{6}(3^2) + \frac{1}{6}(4^2) + \frac{1}{6}(5^2) + \frac{1}{6}(6^2) = \frac{91}{6}.$$

Remark 4.4. Suppose X takes the value $(-2)^k$ with probability 2^{-k} for every positive integer k. Then $|X|$ takes the value 2^k with probability 2^{-k} for every positive integer k. So, $\mathbf{E}\,|X| = \sum_{k\geq 1} 1 = \infty$. So, $\mathbf{E}(X)$ is undefined.

Example 4.5. In a recent Powerball lottery, one ticket costs \$2, and the jackpot was around $\$(1/2)10^9$ (after deducting taxes). The number of people winning the jackpot shares the jackpot. Let X be your profit from buying one lottery ticket. Consider the following simplified version of the lottery. Suppose you either are the only winner of the jackpot, or you lose. There were around $(1/3)10^9$ tickets sold, and around $(1/3)10^9$ distinct possible ticket numbers. Assume that every ticket is chosen uniformly at random among all possible ticket numbers, and whether or not someone wins or loses is independent of everyone else. Let $p = 3\cdot 10^{-9}$. Then the probability that you win and everyone else loses is $p(1-p)^{1/p} \approx p/e \approx p/3$. That is, $\mathbf{P}(X = -2) \approx 1 - p/3$ and $\mathbf{P}(X = (1/2)10^9 - 2) \approx p/3$. So,

$$\mathbf{E}X = -2(1-p) + (1/2)10^9 p \approx -2 + 3/2 = -.5.$$

Since the expected value is negative, it was not sensible to buy a lottery ticket. Also, let N be the number of people who get the winning number. Using the Poisson Approximation

to the Binomial with $\lambda = 1$, we have $\mathbf{P}(N = k) \approx \frac{1}{ek!}$ for any positive integer k. So, $\mathbf{P}(N = 0) \approx 1/e$, $\mathbf{P}(N = 1) \approx 1/e$, $\mathbf{P}(N = 2) \approx 1/(2e) \approx 1/6$, $\mathbf{P}(N = 3) \approx 1/(6e) \approx 1/18$, and so on. So, having two or three winners is not so unexpected.

Proposition 4.6 (**Expected Value Rule**). *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. Then*

$$\mathbf{E}(f(X)) = \sum_{x \in \mathbb{R}} f(x) p_X(x).$$

In particular, if n is a positive integer, we can choose $f(x) = x^n$ to get

$$\mathbf{E}(X^n) = \sum_{x \in \mathbb{R}} x^n p_X(x).$$

Also, if a, b are constants, we can choose $f(x) = ax + b$ to get

$$\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$$

Proof. From Proposition 3.19, $p_{f(X)}(y) = \sum_{x \in \mathbb{R}\colon\, f(x)=y} p_X(x)$. So,

$$\begin{aligned}\mathbf{E}(f(X)) &= \sum_{y \in \mathbb{R}} y p_{f(X)}(y) = \sum_{y \in \mathbb{R}} \sum_{x \in \mathbb{R}\colon\, f(x)=y} y p_X(x) \\ &= \sum_{y \in \mathbb{R}} \sum_{x \in \mathbb{R}\colon\, f(x)=y} f(x) p_X(x) = \sum_{x \in \mathbb{R}} f(x) p_X(x).\end{aligned}$$

In the last equality, we used Exercise 2.21.

Now, let a, b be constants. Using Proposition 4.6 and then Proposition 3.8,

$$\mathbf{E}(aX + b) = \sum_{x \in \mathbb{R}} (ax + b) p_X(x) = a \sum_{x \in \mathbb{R}} x p_X(x) + b \sum_{x \in \mathbb{R}} p_X(x) = a\mathbf{E}(X) + b.$$

□

Definition 4.7 (**Variance**). Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. We define the **variance** of X, denoted $\mathrm{var}(X)$, by

$$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2.$$

We define the **standard deviation** of X, denoted σ_X, by

$$\sigma_X = \sqrt{\mathrm{var}(X)}.$$

The notation $\mathbf{E}(X - \mathbf{E}(X))^2$ is a shorthand for $\mathbf{E}[(X - \mathbf{E}(X))^2]$.

Proposition 4.8 (**Properties of Variance**). *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. Let a, b be constants. Then*

$$\mathrm{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.$$

Moreover,

$$\mathrm{var}(aX + b) = a^2 \mathrm{var}(X).$$

Proof. Using Proposition 4.6 and then Propositions 4.6 and 3.8,

$$\begin{aligned}\operatorname{var}(X) &= \mathbf{E}(X - (\mathbf{E}(X))^2) = \sum_{x\in\mathbb{R}} (x - \mathbf{E}(X))^2 p_X(x) \\ &= \sum_{x\in\mathbb{R}} x^2 p_X(x) - 2\mathbf{E}(X) \sum_{x\in\mathbb{R}} x p_X(x) + (\mathbf{E}(X))^2 \sum_{x\in\mathbb{R}} p_X(x) \\ &= \mathbf{E}(X^2) - 2\mathbf{E}(X)\mathbf{E}(X) + (\mathbf{E}(X))^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.\end{aligned}$$

From Proposition 4.6, $\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$. So, using Proposition 4.6,

$$\begin{aligned}\operatorname{var}(aX + b) &= \mathbf{E}(aX + b - (a\mathbf{E}(X) + b))^2 = \mathbf{E}(aX - a\mathbf{E}(X))^2 = \mathbf{E}(a^2(X - \mathbf{E}(X))^2) \\ &= a^2\mathbf{E}(X - \mathbf{E}(X))^2 = a^2\operatorname{var}(X).\end{aligned}$$

□

Example 4.9. Returning again to Example 4.3, suppose X takes the values $\{1, 2, 3, 4, 5, 6\}$ each with probability $1/6$. We computed $\mathbf{E}(X) = 7/2$, so

$$\begin{aligned}\operatorname{var}(X) &= \mathbf{E}(X - \mathbf{E}(X))^2 \\ &= \frac{1}{6}(1 - \frac{7}{2})^2 + \frac{1}{6}(2 - \frac{7}{2})^2 + \frac{1}{6}(3 - \frac{7}{2})^2 + \frac{1}{6}(4 - \frac{7}{2})^2 + \frac{1}{6}(5 - \frac{7}{2})^2 + \frac{1}{6}(6 - \frac{7}{2})^2 = \frac{35}{12}.\end{aligned}$$

Alternatively, we computed in Example 4.3 that $\mathbf{E}(X^2) = 91/6$. So, by Proposition 4.8, $\operatorname{var}(X) = 91/6 - (7/2)^2 = 182/12 - 147/12 = 35/12$. Lastly, the standard deviation of X is $\sigma_X = \sqrt{35/12} \approx 1.7078$. So, the value of X is typically in the interval $(\mathbf{E}(X) - \sigma_X, \mathbf{E}(X) + \sigma_X) = (3.5 - 1.7078, 3.5 + 1.7078)$.

Example 4.10. Let X be a Poisson random variable with parameter $\lambda > 0$. Then $p_X(k) = e^{-\lambda}\lambda^k/k!$ when k is a nonnegative integer. We then compute

$$\begin{aligned}\mathbf{E}(X) &= \sum_{k=0}^{\infty} k p_X(k) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.\end{aligned}$$

Exercise 4.11. Let X be a discrete random variable taking a finite number of values. Let $t \in \mathbb{R}$. Consider the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(t) = \mathbf{E}(X - t)^2$. Show that the function f takes its minimum value when $t = \mathbf{E}X$. Moreover, if X takes at least two different values, each with some positive probability, then f is uniquely minimized when $t = \mathbf{E}X$.

Exercise 4.12. Let $0 < p < 1$ and let n be a positive integer. Compute the mean of a binomial random variable with parameter p. Then, compute the mean of a Poisson random variable with parameter $\lambda > 0$.

Exercise 4.13. Let X be a nonnegative random variable on a sample space Ω. Assume that X only takes integer values. Prove that

$$\mathbf{E}(X) = \sum_{n=1}^{\infty} \mathbf{P}(X \geq n).$$

Exercise 4.14. As we will see later in the course, the expectation is very closely related to integrals. This exercise gives a hint toward this relation. Let $\Omega = [0,1]$. Let $\mathbf{P}$ be the probability law on Ω such that $\mathbf{P}([a,b]) = \int_a^b dt = b - a$ whenever $0 \leq a < b \leq 1$. Let n be a positive integer. Let $X \colon \Omega \to \mathbb{R}$ be such that X is constant on any interval of the form $[i/n, (i+1)/n)$, whenever $0 \leq i \leq n-1$. Show that

$$\mathbf{E}(X) = \int_0^1 X(t)dt$$

Now, consider a different probability law, where $\mathbf{P}([a,b]) = \int_a^b \frac{1}{2\sqrt{t}} dt$ whenever $0 \leq a < b \leq 1$. Show that

$$\mathbf{E}(X) = \int_0^1 X(t)\frac{1}{2\sqrt{t}}dt.$$

Exercise 4.15. Let $a_1, \ldots, a_n$ be distinct numbers, representing the quality of n people. Suppose n people arrive to interview for a job, one at a time, in a random order. That is, every possible arrival order of these people is equally likely. For each $i \in \{1, \ldots, n\}$, upon interviewing the i^{th} person, if $a_i > a_j$ for all $1 \leq j < i$, then the i^{th} person is hired. That is, if the person currently being interviewed is better than the previous candidates, she will be hired. What is the expected number of hirings that will be made? (Hint: let $X_i = 1$ if the i^{th} person to arrive is hired, and let $X_i = 0$ otherwise. Consider $\sum_{i=1}^n X_i$.)

4.2. **Joint Mass Function, Covariance.**

Definition 4.16 (**Joint PMF**)**.** Let X, Y be two discrete random variables on a sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Let $x, y \in \mathbb{R}$. Define the **joint probability mass function** of X and Y by

$$p_{X,Y}(x,y) = \mathbf{P}(\{X = x\} \cap \{Y = y\}) = \mathbf{P}(X = x \text{ and } Y = y) = \mathbf{P}(X = x, Y = y).$$

Let A be a subset of $\mathbb{R}^2$. We define

$$\mathbf{P}((X,Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y).$$

Proposition 4.17. *Let X, Y be two discrete random variables on a sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Then for any $x, y \in \mathbb{R}$,*

$$p_X(x) = \sum_{t\in\mathbb{R}} p_{X,Y}(x,t), \qquad p_Y(y) = \sum_{t\in\mathbb{R}} p_{X,Y}(t,y).$$

Proof. We prove the first equality, since the second one is proven similarly. For any $t \in \mathbb{R}$, let A_t be the event that $Y = t$. If $t_1 \neq t_2$, then $A_{t_1} \cap A_{t_2} = \emptyset$. And $\cup_{t\in\mathbb{R}} A_t = \Omega$. So, from Axiom (ii) in Definition 2.22,

$$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\cup_{t\in\mathbb{R}}\{X = x\} \cap \{Y = t\}) = \sum_{t\in\mathbb{R}} \mathbf{P}(X = x, Y = t) = \sum_{t\in\mathbb{R}} p_{X,Y}(x,t).$$

□

Remark 4.18. We refer to p_X as the **marginal** PMF of X, and we refer to p_Y as the marginal PMF of Y, to distinguish these PMFs from the joint PMF $p_{X,Y}$.

Proposition 4.19. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X and Y be discrete random variables on Ω taking a finite number of values. Let c be a constant. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. Then*

$$\mathbf{E}f(X,Y) = \sum_{(x,y)\in\mathbb{R}^2} f(x,y)p_{X,Y}(x,y).$$

Consequently, choosing $f(x,y) = x+y$, or $f(x,y) = cx$ where c is a constant,

$$\mathbf{E}(X+Y) = \mathbf{E}(X) + \mathbf{E}(Y), \qquad \mathbf{E}(cX) = c\mathbf{E}(X).$$

So, in linear algebraic terms, $\mathbf{E}$ is a linear transformation.

Proof. Let $z \in \mathbb{R}$. Then $p_{f(X,Y)}(z) = \mathbf{P}(f(X,Y) = z)$. Let $x, y \in \mathbb{R}$. Let $A_{x,y}$ be the event $\{X = x\} \cap \{Y = Y\}$. If $(x_1, y_1) \neq (x_2, y_2)$, then $A_{x_1,y_1} \cap A_{x_2,y_2} = \emptyset$. And $\cup_{(x,y)\in\mathbb{R}^2} A_{x,y} = \Omega$. So, from Axiom (ii) of Definition 2.22,

$$\begin{aligned}\mathbf{P}(f(X,Y) = z) &= \mathbf{P}(\cup_{(x,y)\in\mathbb{R}^2}\{f(X,Y) = z\} \cap A_{x,y}) \\ &= \sum_{(x,y)\in\mathbb{R}^2} \mathbf{P}(\{f(X,Y) = z\} \cap \{X = x\} \cap \{Y = y\}) = \sum_{(x,y)\in\mathbb{R}^2\colon f(x,y)=z} \mathbf{P}(X = x, Y = y).\end{aligned}$$

Note that $\mathbb{R}^2 = \cup_{z\in\mathbb{R}}\{(x,y) \in \mathbb{R}^2\colon f(x,y) = z\}$, where the union is disjoint. So,

$$\begin{aligned}\mathbf{E}(f(X,Y)) &= \sum_{z\in\mathbb{R}} z p_{f(X,Y)}(z) = \sum_{z\in\mathbb{R}} z \sum_{(x,y)\in\mathbb{R}^2\colon f(x,y)=z} \mathbf{P}(X = x, Y = y) \\ &= \sum_{(x,y)\in\mathbb{R}^2} f(x,y)\mathbf{P}(X = x, Y = y)\end{aligned}$$

The first equality is proven. We now consider $f(x,y) = x + y$. We have

$$\begin{aligned}\mathbf{E}(X+Y) &= \sum_{x\in\mathbb{R}} x \sum_{y\in\mathbb{R}} \mathbf{P}(X = x, Y = y) + \sum_{y\in\mathbb{R}} y \sum_{x\in\mathbb{R}} \mathbf{P}(X = x, Y = y) \\ &= \sum_{x\in\mathbb{R}} x\mathbf{P}(X = x) + \sum_{y\in\mathbb{R}} y\mathbf{P}(Y = y) = \mathbf{E}(X) + \mathbf{E}(Y).\end{aligned}$$

In the last line, we used Proposition 4.17 to get $\sum_{y\in\mathbb{R}} \mathbf{P}(\{X = x\} \cap \{Y = y\}) = \mathbf{P}(X = x)$, and $\sum_{x\in\mathbb{R}} \mathbf{P}(\{X = x\} \cap \{Y = y\}) = \mathbf{P}(Y = y)$. Finally, the equality $\mathbf{E}(cX) = c\mathbf{E}(X)$ was proven in Proposition 4.6. □

Exercise 4.20. Suppose there are ten separate bins. You first randomly place a sphere randomly in one of the bins, where each bin has an equal probability of getting the sphere. Once again, you randomly place another sphere uniformly at random in one of the bins. This process occurs twenty times, so that twenty spheres have been placed in bins. What is the expected number of empty bins at the end?

Exercise 4.21. You want to complete a set of 100 baseball cards. Cards are sold in packs of ten. Assume that each card is equally likely to be contained in any pack of cards. How many packs of cards should you buy in order to get a complete set of cards?

Exercise 4.22. Suppose we are drawing cards out of a standard 52 card deck without replacing them. How many cards should we expect to draw out of the deck before we find (a) a King? (b) a Heart?

Exercise 4.23. Let $f\colon \mathbb{R} \to \mathbb{R}$ be twice differentiable function. Assume that f is convex. That is, $f''(x) \geq 0$, or equivalently, the graph of f lies above any of its tangent lines. That is, for any $x, y \in \mathbb{R}$,

$$f(x) \geq f(y) + f'(y)(x - y).$$

(In Calculus class, you may have referred to these functions as "concave up.") Let X be a discrete random variable. By setting $y = \mathbf{E}(X)$, prove **Jensen's inequality**:

$$\mathbf{E}f(X) \geq f(\mathbf{E}(X)).$$

In particular, choosing $f(x) = x^2$, we have $\mathbf{E}(X^2) \geq (\mathbf{E}(X))^2$.

Definition 4.24 (**Covariance**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X and Y be discrete random variables on Ω taking a finite number of values. We define the **covariance** of X and Y, denoted $\operatorname{cov}(X, Y)$, by

$$\operatorname{cov}(X, Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))).$$

Remark 4.25.

$$\operatorname{cov}(X, X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \operatorname{var}(X).$$

The covariance of X and Y is meant to measure whether or not X and Y are related somehow. We will discuss the meaning of covariance a bit more further below. For now, we make the following observation.

Lemma 4.26. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let $X_1, \ldots, X_n$ be discrete random variables on Ω taking a finite number of values. Then*

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{cov}(X_i, X_j).$$

Proof. From Proposition 4.19, $\mathbf{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbf{E}(X_i)$. So,

$$\begin{aligned}
\operatorname{var}(\sum_{i=1}^{n} X_i) &= \mathbf{E}(\sum_{i=1}^{n} X_i - \mathbf{E}(\sum_{i=1}^{n} X_i))^2 = \mathbf{E}(\sum_{i=1}^{n} (X_i - \mathbf{E}(X_i)))^2 \\
&= \mathbf{E}\left(\sum_{i=1}^{n} (X_i - \mathbf{E}(X_i))^2\right) + 2\mathbf{E}\left(\sum_{1 \leq i < j \leq n} (X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))\right) \\
&= \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{cov}(X_i, X_j).
\end{aligned}$$

□

Exercise 4.27. Let n be a positive integer, and let $0 < p < 1$. Let $\Omega = \{0, 1\}^n$. Any $\omega \in \Omega$ can then be written as $\omega = (\omega_1, \ldots, \omega_n)$ with $\omega_i \in \{0, 1\}$ for each $i \in \{1, \ldots, n\}$. Let $\mathbf{P}$ be the probability law described in Example 2.64. That is, for any $\omega \in \Omega$, we have

$$\mathbf{P}(\omega) = \prod_{i=1}^{n} p^{\omega_i}(1 - p)^{1 - \omega_i} = p^{\sum_{i=1}^n \omega_i}(1 - p)^{n - \sum_{i=1}^n \omega_i}.$$

For each $i \in \{1, \ldots, n\}$, define $X_i\colon \Omega \to \mathbb{R}$ so that $X_i(\omega) = \omega_i$ for any $\omega \in \Omega$. That is, if Ω and $\mathbf{P}$ model the flipping of n distinct biased coins, then $X_i = 1$ when the i^{th} coin is heads, and $X_i = 0$ when the i^{th} coin is tails.

First, show that $\mathbf{P}(\Omega) = 1$. Then, compute the expected value of X_i for each $i \in \{1, \ldots, n\}$. Next, compute the expected value of $Y = \sum_{i=1}^{n} X_i$. Finally, prove that Y is a binomial random variable with parameters n and p.

Exercise 4.28 (**Inclusion-Exclusion Formula**)**.** This Exercise gives an alternate proof of the following identity, which is known as the Inclusion-Exclusion Formula: Let $A_1, \ldots, A_n \subseteq \Omega$. Then:

$$\mathbf{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{1 \le i < j \le n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} \mathbf{P}(A_i \cap A_j \cap A_k) \\ \cdots + (-1)^{n+1} \mathbf{P}(A_1 \cap \cdots \cap A_n).$$

Let Y be a random variable such that $Y = 1$ on $\cup_{i=1}^{n} A_i$, and such that $Y = 0$ otherwise.

- Show that $Y = 1 - \prod_{i=1}^{n} (1 - X_i)$.
- Expand out the product in the previous item, and take the expected value of both sides of the result. Deduce the Inclusion-Exclusion formula.

4.2.1. *More than Two Random Variables.* Our results on the joint PMF can be easily extended to any number of random variables. For example, if $X_1, \ldots, X_n$ are discrete random variables, and if $x_1, \ldots, x_n \in \mathbb{R}$, the joint PMF of $X_1, \ldots, X_n$ is defined as

$$p_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Then

$$p_{X_1}(x_1) = \sum_{x_2, \ldots, x_n \in \mathbb{R}} p_{X_1, \ldots, X_n}(x_1, \ldots, x_n),$$

$$p_{X_1, X_2}(x_1, x_2) = \sum_{x_3, \ldots, x_n \in \mathbb{R}} p_{X_1, \ldots, X_n}(x_1, \ldots, x_n), \qquad \text{etc.}$$

Also, if $f \colon \mathbb{R}^n \to \mathbb{R}$ is a function, we have

$$\mathbf{E} f(X_1, \ldots, X_n) = \sum_{x_1, \ldots, x_n \in \mathbb{R}} f(x_1, \ldots, x_n) p_{X_1, \ldots, X_n}(x_1, \ldots, x_n).$$

4.3. **Conditioning.** When dealing with events A, B, we consider the conditional probability $\mathbf{P}(A|B)$ of A given B. We now also discuss conditioning for random variables. Given a random variable X, we can condition X on a set A, and we can also condition X on another random variable Y.

Definition 4.29 (**Conditioning a Random Variable on a Set**)**.** Let X be a discrete random variable on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$ with $\mathbf{P}(A) > 0$. Then the **random variable** X **conditioned on** A, denoted $X|A$, is a random variable with the following PMF:

$$p_{X|A}(x) = \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)}, \qquad \forall\, x \in \mathbb{R}.$$

It follows from Proposition 2.38 that $\sum_{x \in \mathbb{R}} p_{X|A}(x) = 1$.

Example 4.30. This Example follows Example 2.36. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and let $\mathbf{P}$ be the uniform probability law on Ω. Let $A = \{2, 4, 6\}$. That is, A is the event that the die roll is even. Let $X(x) = x$ for all $x \in \Omega$. Then X is the roll of the fair six-sided die. If $x \in A$, then $\mathbf{P}(\{X = x\} \cap A) = \mathbf{P}(X = x)$, and if $x \notin A$, then $\mathbf{P}(\{X = x\} \cap A) = 0$. So,

$$p_{X|A}(x) = \begin{cases} \frac{1/6}{1/2} = \frac{1}{3} & \text{, if } x \in \{2, 4, 6\} \\ 0 & \text{, otherwise.} \end{cases}.$$

So, if we know that the die roll is even, that is, if we know that A occurs, then $X|A$ takes the values $\{2, 4, 6\}$ each with probability $1/3$. Moreover, $X|A$ does not take any odd values, even though X did.

Definition 4.31 (**Conditioning one Random Variable on another**)**.** Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Let $y \in \mathbb{R}$ with $p_Y(y) > 0$. Then the **random variable** X **conditioned on** $Y = y$, is a random variable with the following PMF:

$$p_{X|Y}(x|y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)}, \qquad \forall\, x \in \mathbb{R}.$$

It follows from Proposition 2.38 that $\sum_{x \in \mathbb{R}} p_{X|Y}(x|y) = 1$.

Remark 4.32. By the definition of $p_{X|Y}$, we have:

$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x|y), \qquad \forall\, x, y \in \mathbb{R} \text{ such that } p_Y(y) > 0.$$

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y|x), \qquad \forall\, x, y \in \mathbb{R} \text{ such that } p_X(x) > 0.$$

So, using Proposition 4.17: for any $x \in \mathbb{R}$,

$$p_X(x) = \sum_{y \in \mathbb{R}} p_{X,Y}(x, y) = \sum_{y \in \mathbb{R}:\, p_Y(y) > 0} p_{X,Y}(x, y) = \sum_{y \in \mathbb{R}:\, p_Y(y) > 0} p_Y(y) p_{X|Y}(x|y).$$

That is, if we average over all possibilities of y for $X|Y$, then we just recover X.

4.3.1. *Conditional Expectation.*

Definition 4.33 (**Conditional Expectation**)**.** Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$ with $\mathbf{P}(A) > 0$. Then the **conditional expectation of** X **given** A, denoted $\mathbf{E}(X|A)$ is

$$\mathbf{E}(X|A) = \sum_{x \in \mathbb{R}} x p_{X|A}(x).$$

If $g \colon \mathbb{R} \to \mathbb{R}$, we define

$$\mathbf{E}(g(X)|A) = \sum_{x \in \mathbb{R}} g(x) p_{X|A}(x).$$

Let $y \in \mathbb{R}$ with $\mathbf{P}(Y = y) > 0$. Then the **conditional expectation of** X **given** $Y = y$, denoted $\mathbf{E}(X|Y = y)$ is

$$\mathbf{E}(X|Y = y) = \sum_{x \in \mathbb{R}} x p_{X|Y}(x|y).$$

The quantity $\mathbf{E}(X)$ can be computed by conditioning on another random variable Y. This procedure is analogous to integrating a two-variable function $f(x,y)$ first in the x variable, and then in the y variable. This two-step integration was very useful to use in calculus, and similarly Theorem 4.34 below is very useful for computing expectations.

Theorem 4.34 (**Total Expectation Theorem**). *Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume X and Y only take a finite number of values. Then*

$$\mathbf{E}(X) = \sum_{y\in\mathbb{R}:\ p_Y(y)>0} p_Y(y)\mathbf{E}(X|Y=y).$$

Let $A_1,\dots,A_n$ be disjoint events in Ω such that $\cup_{i=1}^n A_i = \Omega$. Assume $\mathbf{P}(A_i) > 0$ for all $i \in \{1,\dots,n\}$. Then

$$\mathbf{E}(X) = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}(X|A_i).$$

Proof. Starting with Definition 4.2, then using Remark 4.32,

$$\begin{aligned}\mathbf{E}(X) &= \sum_{x\in\mathbb{R}} x p_X(x) = \sum_{x\in\mathbb{R}} x \sum_{y\in\mathbb{R}:\ p_Y(y)>0} p_Y(y)p_{X|Y}(x|y)\\ &= \sum_{y\in\mathbb{R}:\ p_Y(y)>0} p_Y(y) \sum_{x\in\mathbb{R}} x p_{X|Y}(x|y) = \sum_{y\in\mathbb{R}:\ p_Y(y)>0} p_Y(y)\mathbf{E}(X|Y=y).\end{aligned}$$

In the last line, we used Definition 4.33. To deduce the last part of the Theorem, we let Y be a random variable such that, for every $i \in \{1,\dots,n\}$, we have $Y = i$ with probability $\mathbf{P}(A_i)$. That is, $p_Y(i) = \mathbf{P}(A_i)$. Then

$$\mathbf{E}(X) = \sum_{y\in\mathbb{R}:\ p_Y(y)>0} p_Y(y)\mathbf{E}(X|Y=y) = \sum_{i=1}^n p_Y(i)\mathbf{E}(X|Y=i) = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}(X|A_i).$$

In the last equality, we used $p_{X|Y}(x|i) = p_{X|A_i}(x)$ for all $x \in \mathbb{R}$ and for all $i \in \{1,\dots,n\}$, so $\mathbf{E}(X|Y=i) = \mathbf{E}(X|A_i)$ by Definition 4.33. □

Example 4.35. Let's compute the mean and variance of a geometric random variable X. Recall that we have $0 < p < 1$, and for any positive integer k,

$$p_X(k) = (1-p)^{k-1}p.$$

Let A be the event $\{X = 1\}$ and let B be the event $\{X > 1\}$. Then from Definition 4.29, $p_{X|A}(x) = 1$ when $x = 1$ and $p_{X|A}(x) = 0$ otherwise. So,

$$\mathbf{E}(X|A) = \mathbf{E}(X|X=1) = \sum_{x\in\mathbb{R}} x p_{X|A}(x) = 1\cdot 1 = 1.$$

Note that $\mathbf{P}(B) = \mathbf{P}(A^c) = 1 - \mathbf{P}(A) = 1 - p$. Using Definition 4.29 again, $p_{X|B}(k) = \mathbf{P}(\{X = k\} \cap \{X > 1\})/\mathbf{P}(B) = \mathbf{P}(X = k)/\mathbf{P}(B)$ if $k > 1$ is an integer. So,

$$\begin{aligned}\mathbf{E}(X|B) &= \sum_{x\in\mathbb{R}} x p_{X|B}(x) = \sum_{k=2}^{\infty} k(1-p)^{k-1}p/\mathbf{P}(B) = \sum_{k=2}^{\infty} k(1-p)^{k-2}p \\ &= \sum_{k=2}^{\infty}(k-1+1)(1-p)^{k-2}p = \sum_{k=1}^{\infty} k(1-p)^{k-1}p + \sum_{k=2}^{\infty}(1-p)^{k-2}p = \mathbf{E}(X) + 1.\end{aligned}$$

So, using Theorem 4.34,

$$\mathbf{E}(X) = \mathbf{P}(A)\mathbf{E}(X|A) + \mathbf{P}(B)\mathbf{E}(X|B) = p + (1-p)(\mathbf{E}(X) + 1).$$

Solving for $\mathbf{E}(X)$, we get $\mathbf{E}(X)(1 - (1-p)) = p + (1-p)$. So,

$$\mathbf{E}(X) = \frac{1}{p}.$$

We could have also computed $\mathbf{E}(X|B)$ in the following way. Recall that X is the number of times that are needed to flip a biased coin in order to get a heads. The condition $X > 1$ means exactly that the first flip was tails. So, after the first flip, the expected number of remaining flips is $\mathbf{E}(X)$, so the total expected number of flips given B is $1 + \mathbf{E}(X)$.

Using similar reasoning, we get

$$\mathbf{E}(X^2|X = 1) = 1, \qquad \mathbf{E}(X^2|X > 1) = \mathbf{E}((1+X)^2) = 1 + 2\mathbf{E}(X) + \mathbf{E}(X^2).$$

So, using Theorem 4.34,

$$\mathbf{E}(X^2) = p + (1-p)(1 + 2\mathbf{E}(X) + \mathbf{E}(X^2)).$$

Solving for $\mathbf{E}(X^2)$, we get $\mathbf{E}(X^2)(1 - (1-p)) = p + (1-p)(1 + 2/p)$, so that

$$\mathbf{E}(X^2) = \frac{1 + 2/p - 2}{p} = \frac{2}{p^2} - \frac{1}{p}.$$

Therefore, by Proposition 4.8,

$$\operatorname{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

4.4. **Independence of Random Variables.** Recall that sets A, B are independent when $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. The independence of random variables is a bit more involved than the independence of sets, since we will require many equalities to hold.

Definition 4.36 (**Independence of a Random Variable and a Set**)**.** Let X be a discrete random variable on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$. We say that X **is independent of** A if

$$\mathbf{P}(\{X = x\} \cap A) = \mathbf{P}(X = x)\mathbf{P}(A), \qquad \forall\, x \in \mathbb{R}.$$

That is, $\{X = x\}$ is independent of A, for all $x \in \mathbb{R}$. That is, knowing that A has occurred does not changed our knowledge of any value of X.

If $\mathbf{P}(A) > 0$, then X is independent of A when

$$p_{X|A}(x) = p_X(x), \qquad \forall\, x \in \mathbb{R}.$$

Example 4.37. Let $\Omega = \{0,1\}^2$ and let $\mathbf{P}$ be the uniform probability measure on Ω. Then $\mathbf{P}$ models the toss of two distinct fair coins. For any $\omega = (\omega_1, \omega_2) \in \{0,1\}^2$, define $X(\omega) = \omega_1$. That is, $X = 1$ when the first coin toss is heads (1), and $X = 0$ when the first coin toss is tails (0). Let A be the event that the second coin toss is heads. That is, $A = \{(0,1),(1,1)\}$. We will show that X and A are independent.

$$\mathbf{P}(\{X=1\} \cap A) = \mathbf{P}(\{(1,0),(1,1)\} \cap A) = \mathbf{P}(1,1) = 1/4 = (1/2)(1/2) = \mathbf{P}(X=1)\mathbf{P}(A).$$
$$\mathbf{P}(\{X=0\} \cap A) = \mathbf{P}(\{(0,0),(0,1)\} \cap A) = \mathbf{P}(0,1) = 1/4 = (1/2)(1/2) = \mathbf{P}(X=0)\mathbf{P}(A).$$

Therefore, X and A are independent.

Definition 4.38 (Independence of a Random Variable from another). Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that X **is independent of** Y if

$$\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x)\mathbf{P}(Y = y), \qquad \forall\, x, y \in \mathbb{R}.$$

That is, $\{X = x\}$ is independent of $\{Y = y\}$, for all $x, y \in \mathbb{R}$. That is, knowing the values of Y does not changed our knowledge of any value of X. Written another way,

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \qquad \forall\, x, y \in \mathbb{R}.$$

Equivalently, X and Y are independent if and only if

$$p_{X|Y}(x|y) = p_X(x), \qquad \forall\, x, y \in \mathbb{R} \quad \text{with } p_Y(y) > 0.$$

When two random variables are independent, they satisfy many nice properties. For example,

Theorem 4.39. *Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that X and Y are independent. Assume that X and Y take a finite number of values. Then*

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$$

Proof. Using Proposition 4.19 and the equality $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$,

$$\mathbf{E}(XY) = \sum_{x,y \in \mathbb{R}} xy p_{X,Y}(x,y) = \sum_{x \in \mathbb{R}} x p_X(x) \sum_{y \in \mathbb{R}} y p_Y(y) = \mathbf{E}(X)\mathbf{E}(Y).$$

□

Corollary 4.40. *Let $X_1, \ldots, X_n$ be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that $X_1, \ldots, X_n$ are pairwise independent. That is, X_i and X_j are independent whenever $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Assume that $X_1, \ldots, X_n$ take a finite number of values. Then*

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i).$$

Proof. Let $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Then by Theorem 4.39,

$$\operatorname{cov}(X_i, X_j) = \mathbf{E}((X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))) = \mathbf{E}(X_i X_j) - 2\mathbf{E}(X_i)\mathbf{E}(X_j) + \mathbf{E}(X_i)\mathbf{E}(X_j) = 0.$$

So, Lemma 4.26 concludes the proof. □

Exercise 4.41. Let X, Y, Z be discrete random variables. Let $f(y) = \mathbf{E}(X|Y = y)$ for any $y \in \mathbb{R}$. Then $f\colon \mathbb{R} \to \mathbb{R}$ is a function. In more advanced probability classes, we consider the random variable $f(Y)$, which is denoted by $\mathbf{E}(X|Y)$. Show that $\mathbf{E}(X + Z|Y) = \mathbf{E}(X|Y) + \mathbf{E}(Z|Y)$. Then, show that $\mathbf{E}[\mathbf{E}(X|Y)] = \mathbf{E}(X)$. That is, understanding $\mathbf{E}(X|Y)$ can help us to compute $\mathbf{E}(X)$.

Exercise 4.42. Give an example of two random variables X, Y that are independent. Prove that these random variables are independent.

Give an example of two random variables X, Y that are not independent. Prove that these random variables are not independent.

Finally, find two random variables X, Y such that $\mathbf{E}(XY) \neq \mathbf{E}(X)\mathbf{E}(Y)$.

Exercise 4.43. Is it possible to have a random variable X such that X is independent of X? Either find such a random variable X, or prove that it is impossible to find such a random variable X.

Exercise 4.44. Let $0 < p < 1$. Let n be a positive integer. Let $X_1, \ldots, X_n$ be pairwise independent Bernoulli random variables. Compute the expected value of

$$S_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then, compute the variance of $S_n - \mathbf{E}(S_n)$. Describe in words what this variance computation tells you as $n \to \infty$. Particularly, what does S_n "look like" as $n \to \infty$? (Think about what we found in Example 4.1. Also, consider the following statistical interpretation. Suppose each X_i is the result of some poll of person i, where $i \in \{1, \ldots, n\}$. Suppose that each person's response is a Bernoulli random variable with parameter p, and each person's response is independent of each other person's response. Then S_n is the average of the results of the poll. If $S_n - \mathbf{E}(S_n)$ has small variance, then our poll is very accurate. So, how accurate is the poll as $n \to \infty$? Note that the accuracy of the poll does *not* depend on the size of the population you are sampling from!)

Exercise 4.45. Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that X and Y are independent. Assume that X and Y take a finite number of values. Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be functions. Then

$$\mathbf{E}(f(X)g(Y)) = \mathbf{E}(f(X))\mathbf{E}(g(Y)).$$

4.4.1. *Independence of Multiple Random Variables.*

Definition 4.46 (**Independence of Random Variables**)**.** Let $X_1, \ldots, X_n$ be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that $X_1, \ldots, X_n$ **are independent** if

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Remark 4.47. Suppose $X_1, \ldots, X_n$ are discrete, independent random variables taking a finite number of values. Let $f_1, \ldots, f_n$ be functions from $\mathbb{R}$ to $\mathbb{R}$. Similar to Exercise 4.45 we have

$$\mathbf{E}(\prod_{i=1}^{n} f_i(X_i)) = \prod_{i=1}^{n} \mathbf{E}(f_i(X_i)).$$

In particular,

$$\mathbf{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} \mathbf{E}(X_i).$$

Proposition 4.48. *Let $X_1, \ldots, X_n$ be discrete random variables on a sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Assume that $X_1, \ldots, X_n$ are independent. Then, for any subset S of $\{1, \ldots, n\}$, the random variables $\{X_i\}_{i \in S}$ are independent. In particular, $X_1, \ldots, X_n$ are pairwise independent.*

Proof. By reordering indices and iterating, it suffices to show that $X_1, \ldots, X_{n-1}$ are independent. That is, it suffices to show that

$$\mathbf{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i), \qquad \forall\, x_1, \ldots, x_{n-1} \in \mathbb{R}.$$

For any $x_n \in \mathbb{R}$, let $B_{x_n} = \{X_n = x_n\}$. Then $B_{x_n} \cap B_{y_n} = \emptyset$ if $x_n \neq y_n$, $x_n, y_n \in \mathbb{R}$, and $\cup_{x_n \in \mathbb{R}} B_{x_n} = \Omega$. So, using Axiom (ii) for probability laws in Definition 2.22,

$$\begin{aligned}\mathbf{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) &= \mathbf{P}(\{X_1 = x_1\} \cap \cdots \cap \{X_{n-1} = x_{n-1}\} \cap (\cup_{x_n \in \mathbb{R}} B_{x_n})) \\ &= \sum_{x_n \in \mathbb{R}} \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n). \qquad (*)\end{aligned}$$

Similarly,

$$\begin{aligned}\prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i) &= \mathbf{P}(\cup_{x_n \in \mathbb{R}} B_x) \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i) \\ &= \sum_{x_n \in \mathbb{R}} \mathbf{P}(X_n = x_n) \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i) = \sum_{x_n \in \mathbb{R}} \prod_{i=1}^{n} \mathbf{P}(X_i = x_i). \qquad (**)\end{aligned}$$

So, the quantities $(*)$ and $(**)$ are equal, by assumption. □

Exercise 4.49. Find three random variables X_1, X_2, X_3 such that: X_1 and X_2 are independent; X_1 and X_3 are independent; X_2 and X_3 are independent; but such that X_1, X_2, X_3 are not independent.

Exercise 4.50. Let $0 < p < 1$. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Let $S_n = \sum_{i=1}^{n} X_i$. A moment generating function can help use to compute moments in various ways. Let $t \in \mathbb{R}$ and compute the moment generating function of X_i for each $i \in \{1, \ldots, n\}$. That is, show that

$$\mathbf{E}e^{tX_i} = (1 - p) + pe^t.$$

Then, using the product formula for independent random variables, show that

$$\mathbf{E}e^{tS_n} = [(1 - p) + pe^t]^n.$$

By differentiating the last equality at $t = 0$, and using the power series expansion of the exponential function, compute $\mathbf{E}S_n$ and $\mathbf{E}S_n^2$.

Exercise 4.51. $X_1, \ldots, X_n$ be independent discrete random variables. Show that

$$\mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \leq x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

5. Continuous Random Variables

Up until this point, we have mostly focused on discrete random variables. These random variables take either a finite or countable number of values. However, we are often confronted with a continuous range of possible values. For example, if I throw a dart at a board, then there is a continuous range of places that the dart could land. Or, the price of a stock is (for many purposes) any possible positive real number. We now develop the theory of random variables which take a continuous range of values.

5.1. Continuous Random Variables.

Definition 5.1 (**Probability Density Function, PDF**)**.** A **probability density function** or PDF, is a function $f\colon \mathbb{R} \to [0,\infty)$ such that $\int_{-\infty}^{\infty} f(x)dx = 1$, and such that, for any $-\infty \le a \le b \le \infty$, the integral $\int_a^b f(x)dx$ exists.

Definition 5.2 (**Continuous Random Variable**)**.** A random variable X on a sample space Ω is called **continuous** if there exists a probability density function f_X such that, for any $-\infty \le a \le b \le \infty$, we have

$$\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)dx.$$

We call f_X the **probability density function of** X.

Remark 5.3. Let X be a continuous random variable with density function f_X. Then for any $a \in \mathbb{R}$, $\mathbf{P}(X = a) = \int_a^a f_X(x)dx = 0$. Consequently, for any $-\infty < a \le b < \infty$, we have

$$\mathbf{P}(a \le X \le b) = \mathbf{P}(a \le X < b) = \mathbf{P}(a < X \le b) = \mathbf{P}(a < X < b).$$

Remark 5.4. Let $I_1, I_2, \ldots$ be disjoint intervals in the real line $\mathbb{R}$. Let $B = \cup_{i=1}^{\infty} I_i$. Then from Axiom (ii) of Definition 2.22,

$$\mathbf{P}(X \in B) = \mathbf{P}(X \in \cup_{i=1}^{\infty} I_i) = \sum_{i=1}^{\infty} \mathbf{P}(X \in I_i) = \sum_{i=1}^{\infty} \int_{I_i} f_X(x)dx = \int_B f_X(x)dx.$$

The following Theorem is typically proven in advanced analysis classes.

Theorem 5.5 (**Fundamental Theorem of Calculus**)**.** *Let f_X be a probability density function. Then the function $g(t) = \int_{-\infty}^{t} f_X(x)dx$ is continuous at any $t \in \mathbb{R}$. Also, if f_X is continuous at a point x, then g is differentiable at $t = x$, and $g'(x) = f_X(x)$.*

Example 5.6. Let $\Omega = [0,1]$, and define $f_X\colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. Then $\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 dx = 1$, and $f_X(x) \ge 0$ for all $x \in \mathbb{R}$, so f_X is a probability density function. So, if f_X is the density function of X, and if $a \le b$, we have

$$\mathbf{P}(a \le X \le b) = \int_{\max(0,\min(a,1))}^{\max(0,\min(b,1))} dx = \max(0,\min(b,1)) - \max(0,\min(a,1)).$$

In particular, if $0 \le a < b \le 1$, we have $\mathbf{P}(a \le X \le b) = b - a$. When X has this density function f_X, we say X is **uniformly distributed in** $[0,1]$.

Note that f_X is not a continuous function, but we still say that X is continuous since the function $g(t) = \int_{-\infty}^{t} f_X(x)dx$ is continuous, by the Fundamental Theorem of Calculus. Also,

note that f_X only takes two values, but X can take any value in $[0,1]$. Finally, note that g is not differentiable when $t=0$ or $t=1$, but g is differentiable for any other $t \in \mathbb{R}$.

Example 5.7. Let $\Omega = [c,d]$, with $-\infty < c < d < \infty$ and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = \frac{1}{d-c}$ when $x \in [c,d]$, and $f_X(x) = 0$ otherwise. Then $\int_{-\infty}^{\infty} f_X(x)dx = \int_c^d \frac{1}{d-c}dx = 1$, and $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, so f_X is a probability density function. So, if f_X is the density function of X, and if $-\infty < a \leq b < \infty$, we have

$$\mathbf{P}(a \leq X \leq b) = \frac{1}{d-c} \int_{\max(c,\min(a,d))}^{\max(c,\min(b,d))} dx = \frac{1}{d-c} \left(\max(c,\min(b,d)) - \max(c,\min(a,d))\right).$$

In particular, if $c \leq a < b \leq d$, we have $\mathbf{P}(a \leq X \leq b) = \frac{b-a}{d-c}$. When X has the density function f_X, we say that X is **uniformly distributed in** $[c,d]$.

Example 5.8. Let $\Omega = \mathbb{R}$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for all $x \in \mathbb{R}$. Then $\int_{-\infty}^{\infty} f_X(x)dx = 1$ by Exercise 5.10 below and $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, so f_X is a probability density function. So, if f_X is the density function of X, and if $-\infty \leq a \leq b \leq \infty$,

$$\mathbf{P}(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

We call X the **standard Gaussian** random variable or the **standard normal** random variable. The distribution f_X resembles a "bell curve."

The Gaussian comes up in many applications, and it has a certain "universality" property which is studied in more advanced probability classes. For example, if we make a histogram of test scores for a class with a large number of people, then the scores will look something like the distribution $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. And we can replace "test scores" with many other things, and the histogram will remain essentially the same. This is what is meant by "universality."

In general, we can intuitively think of a distribution function f_X as a histogram for the (random) values that X takes.

Example 5.9. Let $\lambda > 0$. Define $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and $f_X(x) = 0$ otherwise. Let's check that f_X satisfies Definition 5.1.

$$\int_{-\infty}^{\infty} f_X(x)dx = \lambda \int_0^{\infty} e^{-\lambda x} dx = \lambda \lim_{N \to \infty} [-\lambda^{-1}(e^{-\lambda N} - 1)] = 1.$$

A random variable X with this density f_X is called an **exponential random variable with parameter** λ. Exponential random variables can be used to model the expiration time of lightbulbs, or other electronic equipment.

Exercise 5.10. Verify that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$. (Hint: let $T = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. It suffices to show that $T^2 = 1$, since $T > 0$. But $T^2 = 1$ follows from Exercise 2.31, after appropriate manipulation.)

5.1.1. *Expected Value.* How should we define the expected value of a continuous random variable? Let's return to Example 5.6. Let $\Omega = [0,1]$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. Then X is uniformly distributed in

$[0,1]$. Let n be a positive integer. We will try to approximate the expected value of X. Consider the intervals $[0,1/n)$, $[1/n,2/n)$, ..., $[(n-1)/n,1)$. Then, for each $i \in \{1,\ldots,n\}$,

$$\mathbf{P}(X \in [(i-1)/n, i/n)) = \int_{(i-1)/n}^{i/n} dx = 1/n.$$

So, to estimate the expected value of X, let's just make the approximation that X takes the value i/n with probability $1/n$, for each $i \in \{1,\ldots,n\}$. This is not quite true, but it is also not so far from the truth. Then we estimate the expected value of X by summing up the (approximate) values of X, multiplied by their probabilities of occurring:

$$\sum_{i=1}^{n} \frac{i}{n} \cdot \mathbf{P}(X \in [(i-1)/n, i/n)) = \sum_{i=1}^{n} \frac{i}{n}\frac{1}{n}.$$

We could compute this sum exactly, but it is perhaps better to see that this sum is a Riemann sum for the function $g(x) = x$ on the interval $[0,1]$. That is,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{i}{n} \cdot \mathbf{P}(X \in [(i-1)/n, i/n)) = \int_0^1 x dx = \int_{-\infty}^{\infty} x f_X(x) dx.$$

The last expression is exactly our definition of expected value for continuous random variables.

Definition 5.11 (**Expected Value**)**.** Let X be a continuous random variable with density function f_X. Assume that $\int_{-\infty}^{\infty} |x|\, f_X(x)dx < \infty$. We then define the **expected value of** X, denoted $\mathbf{E}(X)$, by

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Let $g\colon \mathbb{R} \to \mathbb{R}$ be a function. We define

$$\mathbf{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

In particular, if n is a positive integer, we have

$$\mathbf{E}(X^n) = \int_{-\infty}^{\infty} x^n f_X(x) dx.$$

Comparing Definition 4.2 to Definition 5.11, we see that we have essentially replaced the sums with integrals. Also, we can use the same definition of variance as before.

Definition 5.12 (**Variance**)**.** Let X be a continuous random variable with density function f_X. We define the **variance of** X, denoted $\mathrm{var}(X)$, by

$$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2.$$

Many facts for discrete random variables also apply to continuous random variables. For example, the following restatements of Propositions 4.6 and 4.8 hold, with the same proof as before, where we replace the sums by integrals.

Proposition 5.13 (**Properties of Expected Value**)**.** *Let X be a continuous random variable. Let a, b be constants. Then*

$$\mathbf{E}(aX + b) = a\mathbf{E}(X) + b.$$

Proof. Using Definition 5.11 and Definition 5.1

$$\mathbf{E}(aX+b) = \int_{-\infty}^{\infty} (ax+b) f_X(x) dx = a \int_{-\infty}^{\infty} x f_X(x) + b \int_{-\infty}^{\infty} f_X(x) dx = a\mathbf{E}(X) + b \cdot 1.$$

□

Proposition 5.14 (**Properties of Variance**). *Let X be a continuous random variable. Let a, b be constants. Then*

$$\mathrm{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.$$

Moreover,

$$\mathrm{var}(aX+b) = a^2 \mathrm{var}(X).$$

Proof. Using Definition 5.11 and Definition 5.1,

$$\begin{aligned}
\mathrm{var}(X) &= \mathbf{E}(X - \mathbf{E}(X))^2 = \int_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f_X(x) dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x) dx - 2\mathbf{E}(X) \int_{-\infty}^{\infty} x f_X(x) dx + (\mathbf{E}(X))^2 \int_{-\infty}^{\infty} f_X(x) dx \\
&= \mathbf{E}(X^2) - 2\mathbf{E}(X)\mathbf{E}(X) + (\mathbf{E}(X))^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.
\end{aligned}$$

From Proposition 5.13, $\mathbf{E}(aX+b) = a\mathbf{E}(X) + b$. So, using Definition 5.11,

$$\begin{aligned}
\mathrm{var}(aX+b) &= \mathbf{E}(aX + b - (a\mathbf{E}(X) + b))^2 = \mathbf{E}(aX - a\mathbf{E}(X))^2 = \mathbf{E}(a^2(X - \mathbf{E}(X))^2) \\
&= a^2 \mathbf{E}(X - \mathbf{E}(X))^2 = a^2 \mathrm{var}(X).
\end{aligned}$$

□

Example 5.15. We revisit Example 5.6. Let $\Omega = [0,1]$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. Then X is uniformly distributed in $[0,1]$. We compute

$$\mathbf{E}(X) = \int_0^1 x dx = \frac{1}{2}, \qquad \mathbf{E}(X^2) = \int_0^1 x^2 dx = \frac{1}{3}.$$

$$\mathrm{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

In particular, if X is uniformly distributed in $[0,1]$, then the average value of X is $1/2$.

Example 5.16. We revisit Example 5.8. Let $\Omega = \mathbb{R}$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for all $x \in \mathbb{R}$. Then X is a standard Gaussian random variable. We compute

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \int_0^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} - \int_0^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = 0.$$

Exercise 5.17. Let X be a continuous random variable with distribution function $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $\forall\ x \in \mathbb{R}$. Show that $\mathrm{var}(X) = 1$.

Example 5.18. We reconsider Example 5.9. Let $\lambda > 0$. Define $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and $f_X(x) = 0$ otherwise. Then X is an **exponential random variable with parameter** λ. Using integration by parts, we compute

$$\mathbf{E}(X) = \lambda \int_0^\infty x e^{-\lambda x} dx = -\int_0^\infty x \frac{d}{dx} e^{-\lambda x} dx = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

$$\mathbf{E}(X^2) = \lambda \int_0^\infty x^2 e^{-\lambda x} dx = -\int_0^\infty x^2 \frac{d}{dx} e^{-\lambda x} dx = \int_0^\infty 2x \frac{d}{dx} e^{-\lambda x} dx = \frac{2}{\lambda}\mathbf{E}(X) = \frac{2}{\lambda^2}.$$

$$\operatorname{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Exercise 5.19. Let X be a random variable such that $f_X(x) = x$ when $0 \leq x \leq \sqrt{2}$ and $f_X(x) = 0$ otherwise. Compute $\mathbf{E}X^2$ and $\mathbf{E}X^3$.

5.2. **Cumulative Distribution Function (CDF).** Our treatments of discrete and continuous random variables have been similar but different. We had to repeat ourselves several times, and some concepts seem similar but not identical. Thankfully, a unified treatment of both discrete and continuous random variables can be done. This unified treatment comes from examining the probability that a random variable X satisfies $\mathbf{P}(X \leq x)$, for any $x \in \mathbb{R}$.

Definition 5.20 (**Cumulative Distribution Function**)**.** Let X be a random variable. The **cumulative distribution function of** X, denoted F_X, is a function $F_X \colon \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = \mathbf{P}(X \leq x), \qquad \forall\, x \in \mathbb{R}.$$

Remark 5.21. If X is a discrete random variable, then

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{y \in \mathbb{R} \colon\, y \leq x} p_X(y).$$

If X is a continuous random variable with density function f_X, then

$$F_X(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt.$$

Proposition 5.22 (**Properties of the Distribution Function**)**.** *Let X be a random variable. The cumulative distribution function F_X satisfies the following properties:*

- *If $x \leq y$, then $F_X(x) \leq F_X(y)$.*
- $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$.
- *If X is discrete, then F_X is piecewise constant.*
- *If X is continuous, then F_X is continuous.*

Remark 5.23. If X is a continuous random variable with probability density function f_X, and if f_X is continuous at a point $x \in \mathbb{R}$, then Theorem 5.5 implies that $\frac{d}{dx} F_X(x) = f_X(x)$.

Example 5.24. Let X be a uniformly distributed random variable in $[0,1]$. In Example 5.6, we showed that X has the distribution function f_X where $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. So,

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_0^{\max(0,\min(x,1))} dt = \begin{cases} x, & x \in [0,1] \\ 0, & x < 0 \\ 1, & x > 1. \end{cases}$$

Note also that

$$\frac{d}{dx}F_X(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & x < 0 \text{ or } x > 1 \\ \text{undefined}, & x = 0 \text{ or } x = 1 \end{cases}$$

So, the derivative of F_X may not exist at some points, but $\frac{d}{dx}F_X(x) = f_X(x)$ for any $x \in (-\infty, 0) \cup (0,1) \cup (1, \infty)$.

Example 5.25 (**Maximum of Independent Variables**)**.** Let X_1, X_2 be two independent discrete random variable with identical CDFs. That is, $\mathbf{P}(X_1 \leq x) = \mathbf{P}(X_2 \leq x)$ for all $x \in \mathbb{R}$. Define the random variable Y by

$$Y = \max(X_1, X_2).$$

Using Exercise 4.51, for any $x \in \mathbb{R}$, we have

$$\mathbf{P}(Y \leq x) = \mathbf{P}(X_1 \leq x, X_2 \leq x) = \mathbf{P}(X_1 \leq x)\mathbf{P}(X_2 \leq x) = [\mathbf{P}(X_1 \leq x)]^2.$$

That is, the CDF of Y is the square of the CDF of X_1.

More generally, if $X_1, X_2, \ldots, X_n$ are independent, discrete random variable with identical CDFs, and if

$$Y = \max(X_1, \ldots, X_n),$$

then for any $x \in \mathbb{R}$,

$$\mathbf{P}(Y \leq x) = [\mathbf{P}(X_1 \leq x)]^n.$$

We can think of Y as the maximum score on a test with n test takers, or the longest throw of a shot put, etc.

Example 5.26. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $p = 1/2$, so that $\mathbf{P}(X_i = 0) = \mathbf{P}(X_i = 1) = 1/2$ for all $1 \leq i \leq n$. Also,

$$\mathbf{P}(X_1 \leq x) = \begin{cases} 0 & , \text{ if } x < 0 \\ 1/2 & , \text{ if } 0 \leq x < 1 \\ 1 & , \text{ if } x \geq 1 \end{cases}.$$

Let $Y = \max(X_1, \ldots, X_n)$. Then

$$\mathbf{P}(Y \leq x) = [\mathbf{P}(X_1 \leq x)]^n = \begin{cases} 0 & , \text{ if } x < 0 \\ (1/2)^n & , \text{ if } 0 \leq x < 1 \\ 1 & , \text{ if } x \geq 1 \end{cases}.$$

That is, $p_Y(0) = (1/2)^n$ and $p_Y(1) = 1 - (1/2)^n$. That is, Y is a Bernoulli random variable with parameter $1 - (1/2)^n$.

Exercise 5.27. Let X be a random variable such that $X = 1$ with probability 1. Show that X is not a continuous random variable. That is, there does not exist a probability density function f such that $\mathbf{P}(X \leq a) = \int_{-\infty}^{a} f(x)dx$ for all $x \in \mathbb{R}$. (Hint: if X were continuous, then the function $g(t) = \int_{-\infty}^{a} f(x)dx$ would be continuous, by the Fundamental Theorem of Calculus.)

5.3. **Normal Random Variables.**

Definition 5.28 (**Normal Random Variable**)**.** Let $\mu \in \mathbb{R}$, $\sigma > 0$. A continuous random variable X is said to be **normal** or **Gaussian** with mean μ and variance σ^2 if X has the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall\, x \in \mathbb{R}.$$

In particular, a **standard normal** or **standard Gaussian** random variable is defined to be a normal with $\mu = 0$ and $\sigma = 1$.

Exercise 5.29. Verify that a Gaussian random variable X with mean μ and variance σ^2 actually has mean μ and variance σ^2.

Let $a, b \in \mathbb{R}$ with $a \neq 0$. Show that $aX + b$ is a normal random variable with mean $a\mu + b$ and variance $a^2\sigma^2$.

In particular, conclude that $(X - \mu)/\sigma$ is a standard normal.

The Gaussian is probably one of the single most useful random variables within mathematics, and within applications of mathematics. Here is a sample result that shows the usefulness of the Gaussian.

Theorem 5.30 (**De Moivre-Laplace Theorem**)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $1/2$. Recall that X_1 has mean $1/2$ and variance $1/4$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - (1/2)n}{\sqrt{n}\sqrt{1/4}} \leq a \right) = \int_{-\infty}^{a} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

That is, when n is large, the CDF of $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ is roughly the same as that of a standard normal. In particular, if you flip n fair coins, then the number of heads you get should typically be in the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, when n is large.

Remark 5.31. The random variable $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ has mean zero and variance 1, just like the standard Gaussian. So, the normalizations of $X_1 + \cdots + X_n$ we have chosen are sensible. Also, to explain the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, note that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\left(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq X_1 + \cdots + X_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2} \right) \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left(-\frac{\sqrt{n}}{2} \leq X_1 + \cdots + X_n - \frac{n}{2} \leq \frac{\sqrt{n}}{2} \right) \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left(-1 \leq \frac{X_1 + \cdots + X_n - \frac{n}{2}}{\sqrt{n}/2} \leq 1 \right) = \int_{-1}^{1} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} \approx .6827.
\end{aligned}$$

In fact, there is nothing special about the parameter $1/2$ in the above theorem.

Theorem 5.32 (**De Moivre-Laplace Theorem, Second Version**)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Recall that X_1 has mean p and variance $p(1-p)$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - pn}{\sqrt{n}\sqrt{p(1-p)}} \leq a \right) = \int_{-\infty}^{a} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

In fact, there is nothing special about Bernoulli random variables in the above theorem. (See the Central Limit Theorem in 170B.)

Exercise 5.33. Using the De Moivre-Laplace Theorem, estimate the probability that 10^6 coin flips of fair coins will result in more than $501,000$ heads. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2}dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2}dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9987$.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.

5.4. **Joint PDFs.**

Definition 5.34 (**Joint Probability Density Function, Two Variables**). A **joint probability density function (PDF)** for two random variables is a function $f\colon \mathbb{R}^2 \to [0,\infty)$ such that $\iint_{\mathbb{R}^2} f(x,y)dxdy = 1$, and such that, for any $-\infty \le a < b \le \infty$ and $-\infty \le c < d \le \infty$, the integral $\int_{y=c}^{y=d}\int_{x=a}^{x=b} f_{X,Y}(x,y)dxdy$ exists.

Definition 5.35. Let X, Y be two continuous random variables on a sample space Ω. We say that X and Y are **jointly continuous** with **joint PDF** $f_{X,Y}\colon \mathbb{R}^2 \to [0,\infty)$ if, for any subset $A \subseteq \mathbb{R}^2$, we have

$$\mathbf{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy.$$

In particular, choosing $A = [a,b] \times [c,d]$ with $-\infty \le a < b \le \infty$ and $-\infty \le c < d \le \infty$, we have

$$\mathbf{P}(a \le X \le b, c \le Y \le d) = \int_{y=c}^{y=d}\int_{x=a}^{x=b} f_{X,Y}(x,y)dxdy.$$

We define the **marginal PDF** f_X of X by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad \forall\, x \in \mathbb{R}.$$

We define the **marginal PDF** f_Y of Y by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx, \qquad \forall\, y \in \mathbb{R}.$$

Note that

$$\mathbf{P}(c \le Y \le d) = \mathbf{P}(-\infty \le X \le \infty, c \le Y \le d) = \int_{y=c}^{y=d}\int_{x=-\infty}^{x=\infty} f_{X,Y}(x,y)dxdy.$$

Comparing this formula with Definition 5.2, we see that the marginal PDF of Y is exactly the PDF of Y. Similarly, the marginal PDF of X is the PDF of X.

Example 5.36. In Exercise 2.31, we considered $\Omega = \mathbb{R}^2$, and we defined the probability law

$$\mathbf{P}(A) = \frac{1}{2\pi}\iint_A e^{-(x^2+y^2)/2}dxdy, \qquad \forall\, A \subseteq \Omega.$$

Suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y)\in A)=\frac{1}{2\pi}\iint_A e^{-(x^2+y^2)/2}dxdy.$$

That is, we can think of X as the x-coordinate of a randomly thrown dart, and we can think of Y as the y-coordinate of a randomly thrown dart on the infinite dartboard $\mathbb{R}^2$.

In this case, the marginals are both standard Gaussians:

$$f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\int_{-\infty}^{\infty}e^{-y^2/2}\frac{dy}{\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}e^{-x^2/2},\qquad\forall x\in\mathbb{R}.$$
$$f_Y(y)=\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\int_{-\infty}^{\infty}e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}e^{-y^2/2},\qquad\forall y\in\mathbb{R}.$$

That is, if we only keep track of the x-coordinate of the random dart, then this x-coordinate is a standard Gaussian itself. And if we only keep track of the y-coordinate of the random dart, then this y-coordinate is also a standard Gaussian.

Example 5.37 (**Buffon's Needle**)**.** Suppose a needle of length $\ell>0$ is kept parallel to the ground. The needle is dropped onto the ground with a random position and orientation. The ground has a grid of equally spaced horizontal lines, where the distance between two adjacent lines is $d>0$. Suppose $\ell<d$. What is the probability that the needle touches one of the lines? (Since $\ell<d$, the needle can touch at most one line.)

Let x be the distance of the midpoint of the needle from the closest line. Let θ be the acute angle formed by the needle and any horizontal line. The tip of the needle exactly touches the line when $\sin\theta=x/(\ell/2)=2x/\ell$. So, any part of the needle touches some line if and only if $x\le(\ell/2)\sin\theta$. Since the needle has a uniformly random position and orientation, we model X,Θ as random variables with joint distribution uniform on $[0,d/2]\times[0,\pi/2]$. So,

$$f_{X,\Theta}(x,\theta)=\begin{cases}\frac{4}{\pi d}, & x\in[0,d/2]\text{ and }\theta\in[0,\pi/2]\\ 0, & \text{otherwise.}\end{cases}$$

(Note that $\iint_{\mathbb{R}^2}f_{X,\Theta}(x,\theta)dxd\theta=1$.) And the probability that the needle touches one of the lines is

$$\begin{aligned}\iint_{0\le x\le(\ell/2)\sin\theta}f_{X,\Theta}(x,\theta)dxd\theta&=\int_{\theta=0}^{\theta=\pi/2}\int_{x=0}^{x=(\ell/2)\sin\theta}\frac{4}{\pi d}dxd\theta\\&=\frac{2\ell}{\pi d}\int_{\theta=0}^{\theta=\pi/2}\sin\theta d\theta=\frac{2\ell}{\pi d}[-\cos\theta]_{\theta=0}^{\theta=\pi/2}=\frac{2\ell}{\pi d}.\end{aligned}$$

Note that $x\le\ell/2<d/2$ always, so the set $0\le x\le(\ell/2)\sin\theta$ is still contained in the set $x\in[0,d/2]$.

In particular, when $\ell=d$, the probability is $2/\pi$.

Definition 5.38. Let X,Y be random variables with joint PDF $f_{X,Y}$. Let $g\colon\mathbb{R}^2\to\mathbb{R}$. Then

$$\mathbf{E}g(X,Y)=\iint_{\mathbb{R}^2}g(x,y)f_{X,Y}(x,y)dxdy.$$

In particular,

$$\mathbf{E}(XY)=\iint_{\mathbb{R}^2}xyf_{X,Y}(x,y)dxdy.$$

Exercise 5.39. Let X, Y be random variables with joint PDF $f_{X,Y}$. Let $a, b \in \mathbb{R}$. Using Definition 5.38, show that $\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y$.

5.5. **Conditioning.**

Definition 5.40 (**Conditioning a Continuous Random Variable on a Set**)**.** Let X be a continuous random variable on a sample space Ω. Let $A \subseteq \Omega$ with $\mathbf{P}(A) > 0$. The **conditional PDF** $f_{X|A}$ of X given A is defined to be the function $f_{X|A}$ satisfying

$$\mathbf{P}(X \in B \,|A) = \int_B f_{X|A}(x)dx, \qquad \forall\, B \subseteq \mathbb{R}.$$

Example 5.41. Suppose $A' \subseteq \mathbb{R}$ and we condition on X satisfying $X \in A'$. That is, A is the event $A = \{X \in A'\}$. Then, using Definition 2.35,

$$\mathbf{P}(X \in B \,|A) = \mathbf{P}(X \in B \,|X \in A') = \frac{\mathbf{P}(X \in B, X \in A')}{\mathbf{P}(X \in A')} = \frac{\int_{B \cap A'} f_X(x)dx}{\mathbf{P}(X \in A')}.$$

So, using Definition 5.40, in this case we have

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbf{P}(X \in A')}, & x \in A' \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.42. Suppose you go to the bus stop, and the time T between successive arrivals of the bus is an exponential random variable with parameter $\lambda > 0$. Let $t > 0$. Suppose you go to the bus stop and someone says the last bus came t minutes ago. Let A be the event that $T > t$. That is, we will take it as given that $T > t$, i.e. that up to time t, the bus has not yet arrived. Let X be the time you need to wait until the next bus arrives. Let $x > 0$. Using Definition 2.35 and Example 5.9,

$$\mathbf{P}(X > x|A) = \mathbf{P}(T > t + x|T > t) = \frac{\mathbf{P}(T > t + x, T > t)}{\mathbf{P}(T > t)} = \frac{\mathbf{P}(T > t + x)}{\mathbf{P}(T > t)}$$
$$= \frac{\lambda \int_{t+x}^{\infty} e^{-\lambda s} ds}{\lambda \int_t^{\infty} e^{-\lambda s} ds} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x} = \lambda \int_x^{\infty} e^{-\lambda s} ds.$$

From Definition 5.40, $\mathbf{P}(X > x|A) = \int_x^{\infty} f_{X|A}(x)dx$. That is, $f_{X|A}(x) = \lambda e^{-\lambda x}$. That is, $X|A$ is also an exponential random variable with parameter λ. That is, even though we know the bus has not arrived for t minutes, this does not at all affect our prediction for the arrival of the next bus.

This property is called the **memoryless** property of the exponential random variable.

Definition 5.43. Let X be a continuous random variable on a sample space Ω. Let $A \subseteq \Omega$ with $\mathbf{P}(A) > 0$. Let $f_{X|A}$ be the conditional PDF of X given A. We define the expectation of X given A by

$$\mathbf{E}(X|A) = \int_{-\infty}^{\infty} x f_{X|A}(x)dx.$$

Exercise 5.44. Suppose you go to the bus stop, and the time T between successive arrivals of the bus is anything between 0 and 30 minutes, with all arrival times being equally likely.

Suppose you get to the bus stop, and the bus just leaves as you arrive. How long should you expect to wait for the next bus? What is the probability that you will have to wait at least 15 minutes for the next bus to arrive?

On a different day, suppose you go to the bus stop and someone says the last bus came 10 minutes ago. How long should you expect to wait for the next bus? What is the probability that you will have to wait at least 10 minutes for the next bus to arrive?

We will now investigate versions of the Total Expectation Theorem 4.34 for continuous random variables.

Theorem 5.45. *Let X be a continuous random variable on a sample space Ω. Let $A_1, \ldots, A_n$ be disjoint events in Ω with $\mathbf{P}(A_i) > 0$ for each $i \in \{1, \ldots, n\}$ and $\cup_{i=1}^n A_i = \Omega$. Assume that $f_X, f_{X|A_1}, \ldots, f_{X|A_n}$ are all continuous functions. Then*

$$\mathbf{E}X = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}(X|A_i).$$

Proof. Let $x \in \mathbb{R}$. From Theorem 2.45,

$$\mathbf{P}(X \leq x) = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{P}(X \leq x|A_i).$$

Written another way,

$$\int_{-\infty}^x f_X(t)dt = \sum_{i=1}^n \mathbf{P}(A_i) \int_{-\infty}^x f_{X|A_i}(t)dt.$$

Differentiating in x and applying Theorem 5.5,

$$f_X(x) = \sum_{i=1}^n \mathbf{P}(A_i) f_{X|A_i}(x).$$

Multipling both sides by x and integrating from $-\infty$ to ∞ then completes the proof. □

Exercise 5.46. Let $A_1, A_2, \ldots$ be disjoint events such that $\mathbf{P}(A_i) = 2^{-i}$ for each $i \geq 1$. Assume $\cup_{i=1}^\infty A_i = \Omega$. Let X be a random variable such that $\mathbf{E}(X|A_i) = (-1)^{i+1}$ for each $i \geq 1$. Compute $\mathbf{E}X$.

Definition 5.47 (Conditioning one Random Variable on Another). Let X and Y be continuous random variables with joint PDF $f_{X,Y}$. Fix some $y \in \mathbb{R}$ with $f_Y(y) > 0$. For any $x \in \mathbb{R}$, define the **conditional PDF** of X, given that $Y = y$ by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \qquad \forall\, x \in \mathbb{R}.$$

We also define the **conditional expectation** of X given $Y = y$ by

$$\mathbf{E}(X|Y = y) = \int_{-\infty}^\infty x f_{X|Y}(x|y)dx.$$

From Definition 5.35, note that $\int_{-\infty}^\infty f_{X|Y}(x|y)dx = 1$. So, $f_{X|Y}(x|y)$ is a probability distribution function.

Example 5.48. We continue the dart board example from Exercise 2.31. We suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y) \in A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2}dxdy, \qquad \forall\, A \subseteq \mathbb{R}^2.$$

We verified the marginals are both standard Gaussians:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \forall\, x \in \mathbb{R}, \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \qquad \forall\, y \in \mathbb{R}.$$

So, in this particular example, we have

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{2\pi} e^{-(x^2+y^2)/2}}{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

That is, in this particular example, conditioning X on Y does not at all change X.

Example 5.49. Suppose X and Y have a joint PDF given by $f_{X,Y}(x,y) = \frac{1}{\pi}$ if $x^2 + y^2 \leq 1$, and $f_{X,Y}(x,y) = 0$ otherwise. Let's compute the marginals first, and then determine the conditional PDFs. Let $x, y \in \mathbb{R}$ with $x^2 + y^2 \leq 1$. Using Definition 5.35,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{y=-\sqrt{1-x^2}}^{y=\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2\sqrt{1-x^2}}{\pi}.$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_{x=-\sqrt{1-y^2}}^{x=\sqrt{1-y^2}} \frac{1}{\pi} dx = \frac{2\sqrt{1-y^2}}{\pi}.$$

So, if $x^2 + y^2 \leq 1$, then

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1/\pi}{2\sqrt{1-y^2}/\pi} = \frac{1}{2\sqrt{1-y^2}}.$$

Similarly,

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{1-x^2}}.$$

That is, in this particular example, conditioning X on Y can drastically change X. For example, X conditioned on $Y = 0$, and X conditioned on $Y = 1/2$ have very different PDFs.

The following Theorem is a version of Theorem 4.34 for continuous random variables.

Theorem 5.50 (Total Expectation Theorem). *Let X, Y be continuous random variables. Assume that $f_{X,Y} \colon \mathbb{R}^2 \to \mathbb{R}$ is a continuous function. Then*

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} \mathbf{E}(X|Y=y) f_Y(y) dy.$$

Proof. Using Definitions 5.43 and 5.47, and then Definition 5.35,

$$\begin{aligned}\int_{-\infty}^{\infty} \mathbf{E}(X|Y=y) f_Y(y) dy &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dy dx = \int_{-\infty}^{\infty} x f_X(x) dx = \mathbf{E}X.\end{aligned}$$

□

In the above proof, we used the following Theorem from analysis.

Theorem 5.51 (**Fubini Theorem**). *Let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be a continuous function such that $\iint_{\mathbb{R}^2} |h(x,y)|\,dxdy < \infty$. Then*

$$\iint_{\mathbb{R}^2} h(x,y)dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y)dx \right) dy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y)dy \right) dx.$$

Exercise 5.52. Let X, Y be random variables. For any $y \in \mathbb{R}$, assume that $\mathbf{E}(X|Y=y) = e^{-|y|}$. Also, assume that Y has an exponential distribution with parameter $\lambda = 2$. Compute $\mathbf{E}X$.

5.6. **Independence.**

Definition 5.53. Let X, Y be random variables with joint PDF $f_{X,Y}$. We say that X and Y are **independent** if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \qquad \forall\, x, y \in \mathbb{R}.$$

Equivalently, using Definition 5.47, the random variables X and Y are independent if

$$f_{X|Y}(x|y) = f_X(x), \qquad \forall\, x \in \mathbb{R} \text{ with } f_Y(y) > 0.$$

More generally, random variables $X_1, \ldots, X_n$ with joint PDF $f_{X_1,\ldots,X_n}$ are **independent** if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Example 5.54. We continue Example 5.48. We suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y) \in A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dxdy, \qquad \forall\, A \subseteq \mathbb{R}^2.$$

We showed in Example 5.36 that X and Y are both standard normals. We verified in Example 5.48 that $f_{X|Y}(x|y) = f_X(x)$ forall $x, y \in \mathbb{R}$. So, X and Y are independent.

Proposition 5.55. *Let X, Y be two independent random variables with joint PDF $f_{X,Y}$. Let $A, B \subseteq \mathbb{R}$. Then the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.*

Proof. Using Definition 5.53 and Theorem 5.51,

$$\begin{aligned} \mathbf{P}(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x,y)dydx = \int_A \int_B f_Y(y)dy f_X(x)dx \\ &= (\int_A f_X(x)dx)(\int_B f_Y(y)dy) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B). \end{aligned}$$

□

Theorem 5.56. *Let X, Y be two independent random variables with joint PDF $f_{X,Y}$. Then*

$$\mathbf{E}(XY) = (\mathbf{E}X)(\mathbf{E}Y).$$

More generally, if $g, h\colon \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}(g(X)h(Y)) = (\mathbf{E}g(X))(\mathbf{E}h(Y)).$$

More generally, if $X_1, \ldots, X_n$ are independent random variables with joint PDF $f_{X_1,\ldots,X_n}$, and if $g_1, \ldots, g_n\colon \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}(\prod_{i=1}^{n} g_i(X_i)) = \prod_{i=1}^{n} \mathbf{E}(g_i(X_i)).$$

Proof. We prove the second statement since it implies the first. Using Definitions 5.38 and 5.53, and Theorem 5.51

$$E(g(X)h(Y)) = \iint_{\mathbb{R}^2} g(x)h(y)f_{X,Y}(x,y)dxdy = \iint_{\mathbb{R}^2} g(x)h(y)f_X(x)f_Y(y)dxdy$$
$$= (\int_{\mathbb{R}} g(x)f_X(x)dx)(\int_{\mathbb{R}} h(y)f_Y(y)dy) = (\mathbf{E}g(X))(\mathbf{E}h(Y)).$$

□

Exercise 5.57. Let X, Y be independent random variables with joint PDF $f_{X,Y}$. Show that

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y).$$

Exercise 5.58. Let X and Y be uniformly distributed random variables on $[0,1]$. Assume that X and Y are independent. Compute the following probabilities:

- $\mathbf{P}(X > 3/4)$
- $\mathbf{P}(Y < X)$
- $\mathbf{P}(X + Y < 1/2)$
- $\mathbf{P}(\max(X,Y) > 1/2)$
- $\mathbf{P}(XY < 1/3)$.

Exercise 5.59. Let X_1, Y_1 be random variables with joint PDF f_{X_1,Y_1}. Let X_2, Y_2 be random variables with joint PDF f_{X_2,Y_2}. Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ and let $S\colon \mathbb{R}^2 \to \mathbb{R}^2$ so that $ST(x,y) = (x,y)$ and $TS(x,y) = (x,y)$ for every $(x,y) \in \mathbb{R}^2$. Let $J(x,y)$ denote the determinant of the Jacobian of S at (x,y). Using the change of variables formula from multivariable calculus, show that

$$f_{X_2,Y_2}(x,y) = f_{X_1,Y_1}(S(x,y))\,|J(x,y)|\,.$$

Exercise 5.60 (Numerical Integration). In computer graphics in video games, etc., various integrations are performed in order to simulate lighting effects. Here is a way to use random sampling to integrate a function in order to quickly and accurately render lighting effects. Let $\Omega = [0,1]$, and let $\mathbf{P}$ be the uniform probably law on Ω, so that if $0 \le a < b \le 1$, we have $\mathbf{P}([a,b]) = b - a$. Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbf{P}(X_i \in [a,b]) = b - a$ for all $0 \le a < b \le 1$, for all $i \in \{1, \ldots, n\}$. Let $f\colon [0,1] \to \mathbb{R}$ be a continuous function we would like to integrate. Instead of integrating f directly, we instead compute the quantity

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Show that

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \int_0^1 f(t)dt.$$

$$\lim_{n\to\infty} \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = 0.$$

That is, as n becomes large, $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ is a good estimate for $\int_0^1 f(t)dt$.

5.7. **Joint CDF.**

Definition 5.61 (**Joint CDF**)**.** Let X, Y be random variables. We define the **joint CDF** of X, Y to be the function

$$F_{X,Y}(x,y) = \mathbf{P}(X \leq x, Y \leq y), \quad \forall\, x, y \in \mathbb{R}.$$

More generally, if $X_1, \ldots, X_n$ are random variables, we define the **joint CDF** of $X_1, \ldots, X_n$ to be the function

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n), \quad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Remark 5.62. If X, Y are independent random variables with joint PDF $f_{X,Y}$, then Proposition 5.55 says that

$$F_{X,Y}(x,y) = \mathbf{P}(X \leq x, Y \leq y) = \mathbf{P}(X \leq x)\mathbf{P}(Y \leq y) = F_X(x)F_Y(y).$$

More generally, if $X_1, \ldots, X_n$ are independent random variables with joint PDF $f_{X_1,\ldots,X_n}$, then

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \quad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Remark 5.63. In fact, we can use the last equality as a *definition* in order to define independence of general random variables. That is, we say random variables $X_1, \ldots, X_n$ are independent if

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \quad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

6. Limit Theorem Preliminaries: Covariance, Transforms, Convolution

6.1. **Introduction to Limit Theorems.** Suppose I flip a fair coin 10^9 times. Then I should expect to get roughly $\frac{1}{2}10^9$ heads and $\frac{1}{2}10^9$ tails. This is formalized in the Law of Large Numbers. Or, suppose I have a casino game where the casino wins 51% of the time. Then over a long period of time, the casino will make money; the Law of Large Numbers guarantees that! However, if I do flip 10^9 fair coins, it is unlikely that I will get *exactly* $\frac{1}{2}10^9$ heads. (What is the exact probability?) There will typically be some small fluctuations around $\frac{1}{2}10^9$. But about how close to $\frac{1}{2}10^9$ will the number of heads be? This question is answered precisely by the Central Limit Theorem. In your previous probability class, you may have mentioned the Central Limit Theorem applied to coin flips, which is known as the De Moivre-Laplace Theorem:

Theorem 6.1 (**De Moivre-Laplace Theorem**)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $1/2$, so that $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = 0) = 1/2$. Recall that X_1 has mean $1/2$ and variance $1/4$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - (1/2)n}{\sqrt{n}\sqrt{1/4}} \leq a\right) = \int_{-\infty}^{a} e^{-t^2/2}\frac{dt}{\sqrt{2\pi}}.$$

That is, when n is large, the CDF of $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ is roughly the same as that of a standard normal. In particular, if you flip n fair coins, then the number of heads you get should typically be in the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, when n is large.

Remark 6.2. The random variable $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ has mean zero and variance 1, just like the standard Gaussian. So, the normalizations of $X_1+\cdots+X_n$ we have chosen are sensible. Also, to explain the interval $(n/2-\sqrt{n}/2, n/2+\sqrt{n}/2)$, note that

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}\left(\frac{n}{2}-\frac{\sqrt{n}}{2} \le X_1+\cdots+X_n \le \frac{n}{2}+\frac{\sqrt{n}}{2}\right)\\ = \lim_{n\to\infty} \mathbf{P}\left(-\frac{\sqrt{n}}{2} \le X_1+\cdots+X_n-\frac{n}{2} \le \frac{\sqrt{n}}{2}\right)\\ = \lim_{n\to\infty} \mathbf{P}\left(-1 \le \frac{X_1+\cdots+X_n-\frac{n}{2}}{\sqrt{n}/2} \le 1\right) = \int_{-1}^{1} e^{-t^2/2}\frac{dt}{\sqrt{2\pi}} \approx .6827.\end{aligned}$$

Exercise 6.3. Using the De Moivre-Laplace Theorem, estimate the probability that 1000000 coin flips of fair coins will result in more than 501,000 heads. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2}dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2}dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9987$.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.

Exercise 6.4. Let X and Y be nonnegative random variables. Recall that we can define

$$\mathbf{E}X := \int_0^\infty \mathbf{P}(X > t)dt.$$

Assume that $X \le Y$. Conclude that $\mathbf{E}X \le \mathbf{E}Y$.

More generally, if X satisfies $\mathbf{E}|X| < \infty$, we define $\mathbf{E}X := \mathbf{E}\max(X,0) - \mathbf{E}\max(-X,0)$. If X, Y are any random variables with $X \le Y$, $\mathbf{E}|X| < \infty$ and $\mathbf{E}|Y| < \infty$, show that $\mathbf{E}X \le \mathbf{E}Y$.

6.2. **Continuity of Probability Laws.** Recall that a probability law $\mathbf{P}$ on a sample space Ω satisfies the following three axioms.

(i) For any $A \subseteq \Omega$, we have $\mathbf{P}(A) \ge 0$. (**Nonnegativity**)

(ii) For any $A, B \subseteq \Omega$ such that $A \cap B = \emptyset$, we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

If $A_1, A_2, \ldots \subseteq \Omega$ and $A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \ne j$, then

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k). \qquad (\textbf{Additivity})$$

(iii) We have $\mathbf{P}(\Omega) = 1$. (**Normalization**)

Recall that $\cup_{k=1}^{\infty} A_k = \{x \in \Omega\colon \exists\, k \ge 1,\ x \in A_k\}$ and $\cap_{k=1}^{\infty} A_k = \{x \in \Omega\colon \forall\, k \ge 1,\ x \in A_k\}$.

In this course, we will make several limiting statements about probabilities. For this reason, the following property of probability laws will be quite useful.

Proposition 6.5 (**Continuity of a Probability Law**). *Let* $\mathbf{P}$ *be a probability law on a sample space* Ω. *Let* $A_1, A_2, \ldots$ *be sets in* Ω *which are increasing, so that* $A_1 \subseteq A_2 \subseteq \cdots$. *Then*

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(\cup_{n=1}^{\infty} A_n).$$

In particular, the limit on the left exists.

Proof. First, recall that $A \setminus B := A \cap B^c$ where $A, B \subseteq \Omega$. Now, let $B_1 := A_1$, let $B_2 := A_2 \setminus A_1$, and for any $n \geq 1$, inductively define $B_n := A_n \setminus A_{n-1}$. We claim that $B_1, B_2, \ldots$ are disjoint, and $\cup_{n=1}^k A_n = \cup_{n=1}^k B_n$ for any $1 \leq k \leq \infty$.

To see the first statement, let $i, j \geq 1$ with $i > j$. Since $i - 1 \geq j$, $A_j \subseteq A_{i-1}$, so $A_{i-1}^c \cap A_j = \emptyset$. So

$$B_i \cap B_j = (A_i \setminus A_{i-1}) \cap (A_j \setminus A_{j-1}) = A_i \cap A_{i-1}^c \cap A_j \cap A_{j-1}^c = \emptyset.$$

To see the second statement, let $x \in \cup_{n=1}^k A_n$. Let $m \geq 1$ such that $m = \min\{1 \leq n \leq k \colon x \in A_n\}$. If $m = 1$, then $x \in B_1 = A_1$. If $m > 1$, then $x \notin A_{m-1}$ so $x \in B_m = A_m \setminus A_{m-1}$. So, in any case, $x \in \cup_{n=1}^k B_n$. For the reverse inclusion, let $x \in \cup_{n=1}^k B_n$. Then $x \in B_n$ for some $1 \leq n \leq k$. So $x \in A_n$ since $B_n \subseteq A_n$. So, $x \in \cup_{n=1}^k A_n$. The claim is proven.

Now, using our claim, we have by the second axiom for probability laws,

$$\begin{aligned} \mathbf{P}(\cup_{n=1}^\infty A_n) = \mathbf{P}(\cup_{n=1}^\infty B_n) &= \sum_{n=1}^\infty \mathbf{P}(B_n) = \lim_{k\to\infty} \sum_{n=1}^k \mathbf{P}(B_n) \\ &= \lim_{k\to\infty} \mathbf{P}(\cup_{n=1}^k B_n) = \lim_{k\to\infty} \mathbf{P}(\cup_{n=1}^k A_n) = \lim_{k\to\infty} \mathbf{P}(A_k). \end{aligned}$$

The last line used $A_k \supseteq A_{k-1} \supseteq \cdots \supseteq A_1$. □

A similar statement can be made for a decreasing sequence of sets.

Proposition 6.6 (**Continuity of a Probability Law**). *Let* $\mathbf{P}$ *be a probability law on a sample space* Ω. *Let* $A_1, A_2, \ldots$ *be sets in* Ω *which are decreasing, so that* $A_1 \supseteq A_2 \supseteq \cdots$. *Then*

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(\cap_{n=1}^\infty A_n).$$

Proof. Apply Proposition 6.5 to A_n^c for any $n \geq 1$, and then apply De Morgan's law:

$$\lim_{n\to\infty} \mathbf{P}(A_n) = 1 - \lim_{n\to\infty} \mathbf{P}(A_n^c) = 1 - \mathbf{P}(\cup_{n=1}^\infty A_n^c) = 1 - \mathbf{P}((\cap_{n=1}^\infty A_n)^c) = \mathbf{P}(\cap_{n=1}^\infty A_n).$$

□

Recall that a **random variable** is a function $X \colon \Omega \to \mathbb{R}$.

Definition 6.7 (**Convergence of Real Numbers**). Let $x_1, x_2, \ldots$ be a sequence of real numbers. Let $x \in \mathbb{R}$. We say that $x_1, x_2, \ldots$ **converges** to x if: $\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon)$ such that, for all $n \geq m$, we have $|x_n - x| < \varepsilon$. If $x_1, x_2, \ldots$ converges to x, we denote this by writing

$$x = \lim_{n\to\infty} x_n.$$

Exercise 6.8. Using the definition of convergence, show that the sequence of numbers $1, 1/2, 1/3, 1/4, \ldots$ converges to 0.

Exercise 6.9 (**Uniqueness of limits**). Let $x_1, x_2, \ldots$ be a sequence of real numbers. Let $x, y \in \mathbb{R}$. Assume that $x_1, x_2, \ldots$ converges to x. Assume also that $x_1, x_2, \ldots$ converges to y. Prove that $x = y$. That is, a sequence of real numbers cannot converge to two different real numbers.

6.3. **Derived Distributions.**

Proposition 6.10. *Let X be a continuous random variable with density function $f_X\colon \mathbb{R} \to [0,\infty)$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be continuous. Let $Y := g(X)$. Assume that F_Y is differentiable, where $F_Y(y) = \mathbf{P}(Y \le y)$ for all $y \in \mathbb{R}$. Then for any $y \in \mathbb{R}$,*

$$f_Y(y) = \frac{d}{dy} \int_{\{x \in \mathbb{R}\colon\, g(x) \le y\}} f_X(x)dx.$$

Proof. Let $A \subseteq \mathbb{R}$. Recall that f_X is defined so that

$$\mathbf{P}(X \in A) = \int_A f_X(x)dx.$$

So, if we let $y \in \mathbb{R}$ and if we define $A := \{x \in \mathbb{R}\colon g(x) \le y\}$, we have

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}(g(X) \le y) = \mathbf{P}(X \in A) = \int_A f_X(x)dx = \int_{\{x \in \mathbb{R}\colon\, g(x) \le y\}} f_X(x)dx.$$

So, if F_Y is differentiable, $\frac{d}{dy}F_Y(y) = f_Y(y)$ for all $y \in \mathbb{R}$, completing the proof. □

Example 6.11. Let X be a uniformly distributed random variable on $[-1,1]$, and let $g\colon \mathbb{R} \to \mathbb{R}$ so that $g(x) = x^3$ for any $x \in \mathbb{R}$. Let $Y := g(X)$. Then for any $y \in \mathbb{R}$,

$$f_Y(y) = \frac{d}{dy} \int_{\{x \in \mathbb{R}\colon\, g(x) \le y\}} f_X(x)dx = \frac{d}{dy} \int_{\{x \in [-1,1]\colon\, x^3 \le y\}} \frac{1}{2}dx.$$

If $y < -1$ the integral is zero. If $y > 1$, the integral is 1. And if $y \in [-1,1]$, we have

$$f_Y(y) = \frac{d}{dy}\frac{1}{2}\int_{x=-1}^{x=y^{1/3}} dx = \frac{1}{2}\frac{d}{dy}[y^{1/3}+1] = \frac{1}{6}y^{-2/3}.$$

And if $y \notin [-1,1]$, we have $f_Y(y) = 0$.

Exercise 6.12. Let X be a uniformly distributed random variable on $[-1,1]$. Let $Y := X^2$. Find f_Y.

Exercise 6.13. Let X be a uniformly distributed random variable on $[0,1]$. Let $Y := 4X(1-X)$. Find f_Y.

Example 6.14. Let X be a continuous random variable such that F_X is differentiable. Let $a, b \in \mathbb{R}$ with $a \neq 0$. Let $g(x) := ax + b$ for any $x \in \mathbb{R}$, and let $Y := g(X) = aX + b$. Then for any $y \in \mathbb{R}$, we will show that

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Suppose $a > 0$. Then the function $\mathbf{P}(Y \le y) = \mathbf{P}(aX + b \le y) = \mathbf{P}(X \le (y-b)/a) = F_X((y-b)/a)$ is differentiable with respect to y. So, for any $y \in \mathbb{R}$, the Chain Rule implies

$$f_Y(y) = \frac{d}{dy} \int_{\{x \in \mathbb{R}\colon\, g(x) \le y\}} f_X(x)dx = \frac{d}{dy}F_X((y-b)/a) = \frac{1}{a}f_X((y-b)/a).$$

The case $a < 0$ is demonstrated similarly.

Example 6.15. Let X be a normal random variable with mean μ and variance $\sigma^2 > 0$ where $\sigma > 0$. That is,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall\, x \in \mathbb{R}.$$

Let $a, b \in \mathbb{R}$ with $a > 0$. Let $Y := aX + b$. Then Y is a Gaussian random variable with variance $a^2\sigma^2$ and mean $b + a\mu$:

$$f_Y(y) = \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{(((y-b)/a)-\mu)^2}{2a^2\sigma^2}} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}$$

Definition 6.16 (**Monotonic Function**)**.** Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$. We say that g is **strictly increasing** if, for any $x, y \in I$ with $x > y$, we have $g(x) > g(y)$. We say that g is **strictly decreasing** if, for any $x, y \in I$ with $x > y$, we have $g(x) < g(y)$. We say that g is **strictly monotonic** if g is either strictly increasing or strictly decreasing.

Remark 6.17 (**Strictly Monotonic Functions are Invertible**)**.** Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$ be a strictly monotonic function with range J. As we recall from calculus, g has an inverse. That is, there exists a strictly monotonic function $h\colon J \to I$ such that $g(h(x)) = x$ for every $x \in J$ and $h(g(x)) = x$ for every $x \in I$. Also, as we recall from calculus, if g is differentiable with $g'(x) \neq 0$ for all $x \in I$, then h is differentiable, and by differentiating the identity $h(g(x)) = x$ and applying the chain rule, we get

$$\frac{d}{dx} h(g(x)) = \frac{1}{g'(x)}, \qquad \forall\, x \in I.$$

Or, written another way (defining $y := g(x)$, so that $x = h(y)$),

$$h'(y) = \frac{1}{g'(h(y))}, \qquad \forall\, y \in J.$$

If we graph g and h, then h is obtained by reflecting g across the line $\{(x, y) \in \mathbb{R}^2\colon x = y\}$. Similarly, g is obtained by reflecting h across the line $\{(x, y) \in \mathbb{R}^2\colon x = y\}$.

Proposition 6.18. *Let X be a continuous random variable such that F_X is differentiable. Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$ be a strictly monotonic, differentiable function with range J. Assume that $g'(x) \neq 0$ for every $x \in I$. Let $Y := g(X)$. Let $h\colon J \to I$ be the inverse of g. Then for any $y \in J$,*

$$f_Y(y) = f_X(h(y)) \cdot \left| \frac{d}{dy} h(y) \right| = f_X(h(y)) \cdot \frac{1}{|g'(h(y))|}.$$

Proof. Let $y \in J$. First, assume g is strictly increasing. Then

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}(X \leq h(y)) = F_X(h(y)).$$

Since F_X and h are differentiable, the Chain Rule then proves the first equality. The second equality follows from Remark 6.17, where we noted that

$$\frac{d}{dy} h(y) = \frac{1}{g'(h(y))}, \qquad \forall y \in J.$$

□

Exercise 6.19. Let X be a uniformly distributed random variable on $[0, 1]$. Find the PDF of $-\log(X)$.

Exercise 6.20. Let X be a standard normal random variable. Find the PDF of e^X.

We can perform similar manipulations to find the joint PDF of functions of several random variables.

Example 6.21. Let X, Y be independent exponential random variables with parameter $\lambda = 1$. So, $f_X(x) = e^{-x}$ for any $x \geq 0$ and $f_X(x) = 0$ otherwise. Let $Z := \max(X, Y)$. Then for any $t \in \mathbb{R}$, $\{Z \leq t\} = \{X \leq t, Y \leq t\}$. So, using independence of X, Y,

$$\mathbf{P}(Z \leq t) = \mathbf{P}(X \leq t, Y \leq t) = \mathbf{P}(X \leq t)\mathbf{P}(Y \leq t) = (1 - e^{-t})^2, \qquad \forall\, t \geq 0.$$

So, using the chain rule,

$$f_Z(z) = \frac{d}{dz}\mathbf{P}(Z \leq z) = \begin{cases} 2(1 - e^{-z})e^{-z} & \text{, if } z \geq 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Exercise 6.22. Let X, Y, Z be independent standard Gaussian random variables. Find the PDF of $\max(X, Y, Z)$.

Example 6.23. Let X, Y be independent standard Gaussian random variables. Let $Z := X/|Y|$. For any $t \in \mathbb{R}$, let $A_t := \{(x, y) \in \mathbb{R}^2 \colon x \leq t\,|y|\}$. Then, using polar coordinates, if $t \geq 0$ we have

$$\begin{aligned} \mathbf{P}(Z \leq t) = \mathbf{P}(X \leq t\,|Y|) = \mathbf{P}((X, Y) \in A_t) &= \frac{1}{2\pi}\int_{A_t} e^{-(x^2+y^2)/2} dx dy \\ &= \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=t|y|} e^{-(x^2+y^2)/2} dx dy \\ &= \frac{1}{2\pi}\int_{\theta=\tan^{-1}(1/t)}^{\theta=2\pi-\tan^{-1}(1/t)}\int_{r=0}^{r=\infty} r e^{-r^2/2} dr d\theta \\ &= \frac{1}{2\pi}\int_{\theta=\tan^{-1}(1/t)}^{\theta=2\pi-\tan^{-1}(1/t)} d\theta = 1 - \frac{1}{\pi}\tan^{-1}(1/t). \end{aligned}$$

Similarly, if $t < 0$, then $\mathbf{P}(Z \leq t) = \frac{1}{\pi}\tan^{-1}(1/|t|)$. So, from the Chain rule,

$$f_Z(z) = \frac{1}{\pi(z^2 + 1)}, \qquad \forall\, z \in \mathbb{R}.$$

Exercise 6.24. Let X be a random variable uniformly distributed in $[0, 1]$ and let Y be a random variable uniformly distributed in $[0, 2]$. Suppose X and Y are independent. Find the PDF of X/Y^2.

6.4. **Covariance.** Recall that the covariance of two random variables X and Y, denoted $\operatorname{cov}(X, Y)$, is

$$\operatorname{cov}(X, Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))).$$

In particular, $\operatorname{cov}(X, X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \operatorname{var}(X)$.

Definition 6.25. Let X, Y be random variables. We say that X, Y are **uncorrelated** if $\operatorname{cov}(X, Y) = 0$.

Exercise 6.26. Let X, Y be random variables with $\mathbf{E}X^2 < \infty$ and $\mathbf{E}Y^2 < \infty$. Prove the **Cauchy-Schwarz inequality**:

$$\mathbf{E}(XY) \leq (\mathbf{E}X^2)^{1/2}(\mathbf{E}Y^2)^{1/2}.$$

Then, deduce the following when X, Y both have finite variance:

$$|\mathrm{cov}(X, Y)| \leq (\mathrm{var}(X))^{1/2}(\mathrm{var}(Y))^{1/2}.$$

(Hint: in the case that $\mathbf{E}Y^2 > 0$, expand out the product $\mathbf{E}(X - Y\mathbf{E}(XY)/\mathbf{E}Y^2)^2$.)

Recall in Lemma 4.26, we proved the following for discrete random variables, though the proof applies for any random variables.

Lemma 6.27. *Let $X_1, \ldots, X_n$ be random variables with $\mathrm{var}(X_i) < \infty$ for all $1 \leq i \leq n$. Then*

$$\mathrm{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \mathrm{cov}(X_i, X_j).$$

Proof.

$$\begin{aligned}
\mathrm{var}(\sum_{i=1}^n X_i) &= \mathbf{E}(\sum_{i=1}^n X_i - \mathbf{E}(\sum_{i=1}^n X_i))^2 = \mathbf{E}(\sum_{i=1}^n (X_i - \mathbf{E}(X_i)))^2 \\
&= \mathbf{E}\left(\sum_{i=1}^n (X_i - \mathbf{E}(X_i))^2\right) + 2\mathbf{E}\left(\sum_{1 \leq i < j \leq n} (X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))\right) \\
&= \sum_{i=1}^n \mathrm{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \mathrm{cov}(X_i, X_j).
\end{aligned}$$

The assumption $\mathrm{var}(X_i) < \infty$ for all $1 \leq i \leq n$ and Exercise 6.26 ensure that all of the above quantities are finite. □

As in Corollary 4.40, Lemma 6.27 immediately implies:

Corollary 6.28. *Let $X_1, \ldots, X_n$ be random variables that are pairwise uncorrelated. That is, $\mathrm{cov}(X_i, X_j) = 0$ for any $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Then*

$$\mathrm{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}(X_i).$$

Corollary 6.29. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$\mathrm{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}(X_i).$$

Proof. Let $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Then, using independence,

$$\mathrm{cov}(X_i, X_j) = \mathbf{E}((X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))) = \mathbf{E}(X_iX_j) - 2\mathbf{E}(X_i)\mathbf{E}(X_j) + \mathbf{E}(X_i)\mathbf{E}(X_j) = 0.$$

So, Corollary 6.28 concludes the proof. □

Exercise 6.30. Let X be a binomial random variable with parameters $n = 2$ and $p = 1/2$. So, $\mathbf{P}(X = 0) = 1/4$, $\mathbf{P}(X = 1) = 1/2$ and $\mathbf{P}(X = 2) = 1/4$. And X satisfies $\mathbf{E}X = 1$ and $\mathbf{E}X^2 = 3/2$.

Let Y be a geometric random variable with parameter $1/2$. So, for any positive integer k, $\mathbf{P}(Y = k) = 2^{-k}$. And Y satisfies $\mathbf{E}Y = 2$ and $\mathbf{E}Y^2 = 6$.

Let Z be a Poisson random variable with parameter 1. So, for any nonnegative integer k, $\mathbf{P}(Z = k) = \frac{1}{e}\frac{1}{k!}$. And Z satisfies $\mathbf{E}Z = 1$ and $\mathbf{E}Z^2 = 2$.

Let W be a discrete random variable such that $\mathbf{P}(W = 0) = 1/2$ and $\mathbf{P}(W = 4) = 1/2$, so that $\mathbf{E}W = 2$ and $\mathbf{E}W^2 = 8$.

Assume that X, Y, Z and W are all independent. Compute

$$\operatorname{var}(X + Y + Z + W).$$

Exercise 6.31. Let $X_1, \ldots, X_n$ be random variables with finite variance. Define an $n \times n$ matrix A such that $A_{ij} = \operatorname{cov}(X_i, X_j)$ for any $1 \leq i, j \leq n$. Show that the matrix A is positive semidefinite. That is, show that for any $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, we have

$$y^T A y = \sum_{i,j=1}^{n} y_i y_j A_{ij} \geq 0.$$

6.5. **Transforms.** Generally speaking, a transform is a way of creating one function from another function. For example, the moment generating function associates a real-valued function to a random variable. And the characteristic function (or Fourier transform) associates a complex-valued function to a random variable.

Definition 6.32 (**Moment Generating Function**)**.** Let X be a random variable. The **moment generating function** of X is a function $M_X \colon \mathbb{R} \to \mathbb{R}$ defined by

$$M_X(t) := \mathbf{E}(e^{tX}), \quad \forall\, t \in \mathbb{R}.$$

Remark 6.33. For certain random variables X, the moment generating function may not exist. For example, if X is a continuous random variable with density function $f_X(x) = x^{-2}$ for any $x > 1$, and $f_X(x) = 0$ otherwise. Then $M_X(t) = \int_1^\infty e^{tx} f_X(x) dx$ does not exist when $t > 0$.

Assume that $M_X(t)$ exists for all $t \in \mathbb{R}$, and assume we can differentiate under the expected value. Then

$$\frac{d}{dt}|_{t=0} M_X(t) = \mathbf{E}\left(\frac{d}{dt}_{t=0} e^{tX}\right) = \mathbf{E}(X).$$

That is, the first derivative of the moment generating function at $t = 0$ is equal to the first moment of X. More generally, the n^{th} derivative of the moment generating function at $t = 0$ is equal to the n^{th} moment of X:

Exercise 6.34. Let X be a random variable. Assume that $M_X(t)$ exists for all $t \in \mathbb{R}$, and assume we can differentiate under the expected value any number of times. For any positive integer n, show that

$$\frac{d^n}{dt^n}|_{t=0} M_X(t) = \mathbf{E}(X^n).$$

So, in principle, all moments of X can be computed just by taking derivatives of the moment generating function.

Example 6.35. Let X be an exponential random variable with parameter $\lambda > 0$. That is, $f_X(x) = \lambda e^{-\lambda x}$ for any $x \geq 0$, and $f_X(x) = 0$ otherwise. Then for any $t < \lambda$,

$$M_X(t) = \lambda \int_0^\infty e^{tx} e^{-\lambda x} dx = \lambda \int_0^\infty e^{(t-\lambda)x} dx$$
$$= \lambda \lim_{N \to \infty} \frac{1}{t-\lambda} [e^{(t-\lambda)x}]_{x=0}^{x=N} = \frac{\lambda}{\lambda - t}.$$

From Exercise 6.34, $\mathbf{E}X = \frac{d}{dt}|_{t=0} M_X(t) = \frac{\lambda}{\lambda^2} = \lambda^{-1}$. More generally, it follows by induction that for any integer $n > 0$,

$$\mathbf{E}X^n = \frac{d^n}{dt^n}|_{t=0} M_X(t) = n! \lambda^{-n}.$$

Instead of proving this equality by induction, we use power series. Let $t \in \mathbb{R}$ with $|t| < 1$. From the summation formula for geometric series,

$$\frac{1}{1-t} = \sum_{k=0}^\infty t^n.$$

That is, for any $t \in \mathbb{R}$ with $|t| < \lambda$,

$$M_X(t) = \frac{\lambda}{\lambda - t} = \frac{1}{1 - (t/\lambda)} = \sum_{k=0}^\infty (t/\lambda)^k.$$

So, from Exercise 6.34, if n is a positive integer, then

$$\mathbf{E}X^n = \frac{d^n}{dt^n}|_{t=0} M_X(t) = \sum_{k=0}^\infty \frac{d^n}{dt^n}|_{t=0} (t/\lambda)^k = \frac{d^n}{dt^n}|_{t=0} (t/\lambda)^n = n! \lambda^{-n}.$$

Exercise 6.36. Let X be a standard Gaussian random variable. Compute an explicit formula for the moment generating function of X. (Hint: completing the square might be helpful.) From this explicit formula, compute an explicit formula for all moments of the Gaussian random variable. (The $2n^{th}$ moment of X should be something resembling a factorial.)

Proposition 6.37. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$M_{X_1 + \cdots + X_n}(t) = \prod_{j=1}^n M_{X_j}(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Since $X_1, \ldots, X_n$ are independent, $e^{tX_1}, \ldots, e^{tX_n}$ are independent, for any $t \in \mathbb{R}$. So,

$$M_{X_1 + \cdots + X_n}(t) = \mathbf{E} e^{t(X_1 + \cdots + X_n)} = \mathbf{E} \prod_{j=1}^n e^{tX_j} = \prod_{j=1}^n \mathbf{E} e^{tX_j} = \prod_{j=1}^n M_{X_j}(t)$$

□

Example 6.38. Let X be a binomial distributed random variable with parameters n and $0 < p < 1$. That is, X can be written as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$ with parameter p. Then by Proposition 6.37, for any $t \in \mathbb{R}$,

$$M_X(t) = \prod_{j=1}^n M_{X_j}(t) = (M_{X_1}(t))^n = ((1-p)e^{0 \cdot t} + pe^t)^n = (1 - p + pe^t)^n.$$

In some cases, the moment generating function uniquely determines the random variable.

Theorem 6.39 (**Lévy Continuity Theorem, Weak Form**). *Let X, Y be random variables. Assume that $M_X(t), M_Y(t)$ exist for all $t \in \mathbb{R}$, and that $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$. Then X and Y have the same CDF.*

Exercise 6.40. Construct two random variables $X, Y \colon \Omega \to \mathbb{R}$ such that $X \neq Y$ but $M_X(t), M_Y(t)$ exist for all $t \in \mathbb{R}$, and such that $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$.

Exercise 6.41. Unfortunately, there exist random variables X, Y such that $\mathbf{E}X^n = \mathbf{E}Y^n$ for all $n = 1, 2, 3, \ldots$, but such that X, Y do not have the same CDF. First, explain why this does not contradict the Lévy Continuity Theorem, Weak Form. Now, let $-1 < a < 1$, and define a density

$$f_a(x) := \begin{cases} \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\log x)^2}{2}} (1 + a \sin(2\pi \log x)) & \text{, if } x > 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Suppose X_a has density f_a. If $-1 < a, b < 1$, show that $\mathbf{E}X_a^n = \mathbf{E}X_b^n$ for all $n = 1, 2, 3, \ldots$. (Hint: write out the integrals, and make a change of variables $s = \log(x) - n$.)

From Exercise 6.34, the moment generating function of a random variable X contains all information about the moments of X. However, as mentioned in Remark 6.33, $M_X(t)$ may not exist for many values of t. So, studying the moment generating function may not be so helpful for certain random variables. Fortunately, the closely related characteristic function will always exist, and it also contains all information about the moments of X

Definition 6.42 (**Characteristic Function/ Fourier Transform**). Let $i := \sqrt{-1}$. Let X be a random variable. The **characteristic function** (or **Fourier transform**) of X is the function $\phi_X \colon \mathbb{R} \to \mathbb{C}$ defined by

$$\phi_X(t) := \mathbf{E}(e^{itX}), \quad \forall\, t \in \mathbb{R}.$$

Or equivalently,

$$\phi_X(t) = M_X(it), \quad \forall\, t \in \mathbb{R}.$$

Remark 6.43 (**Expectation of Complex-Valued Random Variables**). Any complex number $z \in \mathbb{C}$ can be written as $z = a + bi$ where $a, b \in \mathbb{R}$. We also define $|z| := \sqrt{a^2 + b^2}$. We call a the real part of z, and we call b the imaginary part of z. Similarly, if Z is a complex-valued random variable, we can write $Z = X + iY$ where X, Y are real-valued random variables. Then, we can define

$$\mathbf{E}Z := \mathbf{E}X + i(\mathbf{E}Y).$$

That is, taking the expected value of a complex-valued random variable is barely different from taking the expected value of a real-valued random variable.

Exercise 6.44. Compute the characteristic function of a uniformly distributed random variable on $[-1, 1]$. (Some of the following formulas might help to simplify your answer: $e^{it} = \cos(t) + i \sin(t)$, $\cos(t) = [e^{it} + e^{-it}]/2$, $\sin(t) = [e^{it} - e^{-it}]/[2i]$, $t \in \mathbb{R}$.)

Remark 6.45. If $t \in \mathbb{R}$, then $|e^{it}| = |\cos(t) + i\sin(t)| = \sqrt{\cos^2(t) + \sin^2(t)} = 1$. The characteristic function is often more appealing to work with than the moment generating function, since the characteristic function always exists. For example, for any $t \in \mathbb{R}$,

$$|\phi_X(t)| = \left|\mathbf{E}e^{itX}\right| \leq \mathbf{E}\left|e^{itX}\right| = 1.$$

However, as mentioned in Remark 6.33, $M_X(t)$ may or may not exist for some $t \in \mathbb{R}$.

Exercise 6.46. Let X be a random variable. Assume we can differentiate under the expected value of $\mathbf{E}e^{itX}$ any number of times. For any positive integer n, show that

$$\frac{d^n}{dt^n}|_{t=0}\phi_X(t) = i^n\mathbf{E}(X^n).$$

So, in principle, all moments of X can be computed just by taking derivatives of the characteristic function.

Exercise 6.47. Let X be a random variable such that $\mathbf{E}|X|^3 < \infty$. Prove that for any $t \in \mathbb{R}$,

$$\mathbf{E}e^{itX} = 1 + it\mathbf{E}X - t^2\mathbf{E}X^2/2 + o(t^2).$$

That is,

$$\lim_{t\to 0} t^{-2}\left|\mathbf{E}e^{itX} - [1 + it\mathbf{E}X - t^2\mathbf{E}X^2/2]\right| = 0$$

(Hint: it may be helpful to use Jensen's inequality, Exercise 4.23, to first justify that $\mathbf{E}|X| < \infty$ and $\mathbf{E}X^2 < \infty$. Then, use the Taylor expansion with error bound: $e^{iy} = 1 + iy - y^2/2 - (i/2)\int_0^y (y-s)^2 e^{is}ds$, which is valid for any $y \in \mathbb{R}$.)

Actually, this same bound holds only assuming $\mathbf{E}X^2 < \infty$, but the proof of that bound requires things we have not discussed.

Since $\phi_X(t) = M_X(it)$, the proof of Proposition 2.23 immediately implies:

Proposition 6.48. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$\phi_{X_1+\cdots+X_n}(t) = \prod_{j=1}^n \phi_{X_j}(t), \qquad \forall\, t \in \mathbb{R}.$$

The Gaussian density has the rather remarkable property that it is essentially its own Fourier transform.

Proposition 6.49. *Let X be a standard Gaussian random variable. Then*

$$\mathbf{E}e^{itX} = e^{-t^2/2}, \qquad \forall\, t \in \mathbb{R}.$$

Proof. Using $e^{itx} = \cos(tx) + i\sin(tx)$ for any $t, x \in \mathbb{R}$,

$$\begin{aligned}
\phi_X(t) = \mathbf{E}e^{itX} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (\cos(tx) + i\sin(tx))e^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \cos(tx)e^{-x^2/2}dx, \qquad \text{since } e^{-x^2/2}\sin(tx) \text{ is odd.}
\end{aligned}$$

Now, differentiating under the integral sign (which is valid, but we will not justify it), and integrating by parts,

$$\begin{aligned}\frac{d}{dt}\phi_X(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(-x)\sin(tx)e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sin(tx)\frac{d}{dx}e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(-t)\cos(tx)e^{-x^2/2}dx = -t\phi_X(t).\end{aligned}$$

Therefore,

$$\frac{d}{dt}[\phi_X(t)e^{t^2/2}] = [t\phi_X(t) - t\phi_X(t)]e^{t^2/2} = 0, \qquad \forall\, t \in \mathbb{R}.$$

That is, there exists a constant $c \in \mathbb{R}$ such that $\phi_X(t)e^{t^2/2} = c$, i.e. $\phi_X(t) = ce^{-t^2/2}$. Since $\phi_X(0) = 1 = c$, the proof is complete. □

6.6. **Sums of Independent Random Variables and Convolution.** Let X, Y be independent random variables. From Proposition 6.37, the moment generating function of $X+Y$ can be easily expressed as $M_{X+Y}(t) = M_X(t)M_Y(t)$, for any t such that both quantities on the right exist. On the other hand, the CDF of $X + Y$ has a more complicated dependence on X and Y.

Example 6.50. Let X, Y be independent integer-valued random variables. Let $t \in \mathbb{Z}$. Then, repeatedly using properties of probability laws, and using that X, Y are independent,

$$\begin{aligned}\mathbf{P}(X+Y=t) &= \sum_{j,k\in\mathbb{Z}:\ j+k=t}\mathbf{P}(X=j, Y=k) = \sum_{j\in\mathbb{Z}}\mathbf{P}(X=j, Y=t-j) \\ &= \sum_{j\in\mathbb{Z}}\mathbf{P}(X=j)\mathbf{P}(Y=t-j) = \sum_{j\in\mathbb{Z}}p_X(j)p_Y(t-j).\end{aligned}$$

Definition 6.51 (**Convolution on the integers**). Let $g, h\colon \mathbb{Z} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h\colon \mathbb{Z} \to \mathbb{R}$ defined by

$$(g*h)(t) := \sum_{j\in\mathbb{Z}} g(j)h(t-j), \qquad \forall\, t \in \mathbb{Z}.$$

So, if X, Y are independent integer-valued random variables, $p_{X+Y}(t) = (p_X * p_Y)(t)\ \forall\, t \in \mathbb{Z}$.

Example 6.52. Let $g(k) := e^{-k}$ and let $h(k) := e^{-k}$ for any nonnegative integer $k \geq 0$, and let $g(k) = h(k) = 0$ for any other integer $k < 0$. Then if $t \geq 0$ is an integer,

$$(g*h)(t) = \sum_{k\in\mathbb{Z}} g(k)h(t-k) = \sum_{k=0}^{t} e^{-k}e^{-(t-k)} = \sum_{k=0}^{t} e^{-t} = (t+1)e^{-t}.$$

And $(g * h)(t) = 0$ for any negative integer t.

A similar formula holds for continuous random variables. That is, if X, Y are two continuous random variables, then the density of $X + Y$ is the convolution of f_X and f_Y.

Definition 6.53 (**Convolution on the real line**). Let $g, h\colon \mathbb{R} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h\colon \mathbb{R} \to \mathbb{R}$ defined by

$$(g*h)(t) := \int_{-\infty}^{\infty} g(x)h(t-x)dx, \qquad \forall\, t \in \mathbb{R}.$$

Proposition 6.54. *Let X, Y be two continuous independent random variables such that $\mathbf{P}(X+Y \le t)$ is differentiable with respect to $t \in \mathbb{R}$. Then*

$$f_{X+Y}(t) = (f_X * f_Y)(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Let X, Y be independent continuous random variables. Then, changing variables,

$$\mathbf{P}(X+Y \le t) = \int_{\{(x,y)\in\mathbb{R}^2 :\, x+y\le t\}} f_{X,Y}(x,y)dxdy = \int_{x=-\infty}^{x=\infty}\int_{y=-\infty}^{y=t-x} f_X(x)f_Y(y)dydx.$$

Then, since $\mathbf{P}(X+Y \le t)$ is differentiable with respect to t, we have by the Fundamental Theorem of Calculus,

$$f_{X+Y}(t) = \frac{d}{dt}\mathbf{P}(X+Y \le t) = \int_{x=-\infty}^{x=\infty} f_X(x)\frac{d}{dt}\int_{y=-\infty}^{y=t-x} f_Y(y)dydx = \int_{x=-\infty}^{x=\infty} f_X(x)f_Y(t-x)dx.$$

□

Example 6.55. Let $g(x) = h(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ for any $x \in \mathbb{R}$. Then if $t \in \mathbb{R}$, we complete the square and change variables twice to get

$$\begin{aligned}(g*h)(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2/2}e^{-(t-x)^2/2}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2+xt-t^2/2}dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(x-t/2)^2+t^2/4-t^2/2}dx = e^{-t^2/4}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(x-t/2)^2}dx \\ &= e^{-t^2/4}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}dx = e^{-t^2/4}\frac{1}{2\sqrt{\pi}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx = e^{-t^2/4}\frac{1}{2\sqrt{\pi}}.\end{aligned}$$

And $(g*h)(t) = e^{-t^2/4}\frac{1}{2\sqrt{\pi}}$ for any $t \in \mathbb{R}$.

Alternatively, we know that if X, Y are independent standard Gaussian random variables, then $X+Y$ is a Gaussian random variable with mean zero and variance $\sigma^2 = 2$. That is, $X+Y$ has density $e^{-t^2/4}\frac{1}{2\sqrt{\pi}}$, $t \in \mathbb{R}$.

Exercise 6.56. Let X, Y, Z be independent and uniformly distributed on $[0,1]$. Note that f_X is not a continuous function.

Using convolution, compute f_{X+Y}. Draw f_{X+Y}. Note that f_{X+Y} is a continuous function, but it is not differentiable at some points.

Using convolution, compute f_{X+Y+Z}. Draw f_{X+Y+Z}. Note that f_{X+Y+Z} is a differentiable function, but it does not have a second derivative at some points.

Make a conjecture about how many derivatives $f_{X_1+\cdots+X_n}$ has, where $X_1, \ldots, X_n$ are independent and uniformly distributed on $[0,1]$. You do not have to prove this conjecture. The idea of this exercise is that convolution is a kind of average of functions. And the more averaging you do, the more derivatives $f_{X_1+\cdots+X_n}$ has.

Exercise 6.57. Construct two random variables X, Y such that X and Y are each uniformly distributed on $[0,1]$, and such that $\mathbf{P}(X+Y=1) = 1$.

Then construct two random variables W, Z such that W and Z are each uniformly distributed on $[0,1]$, and such that $W+Z$ is uniformly distributed on $[0,2]$.

(Hint: there is a way to do each of the above problems with about one line of work. That is, there is a way to solve each problem without working very hard.)

7. Limit Theorems

We now start to build up the machinery that is used to prove the two "big theorems" of probability: the Law of Large Numbers, and the Central Limit Theorem. We begin with some useful inequalities.

7.1. Markov and Chebyshev Inequalities.

Markov's inequality says that a random variable with finite expected value cannot be too large very often.

Proposition 7.1 (**The Markov Inequality**). *Let X be a nonnegative random variable. Then*

$$\mathbf{P}(X \geq t) \leq \frac{\mathbf{E}X}{t}, \qquad \forall\, t > 0.$$

Proof. Let $t > 0$. Let Y be a random variable such that

$$Y = \begin{cases} t & \text{, if } X \geq t \\ 0 & \text{, if } X < t. \end{cases}$$

By definition of Y, we have $Y \leq X$. Therefore, $\mathbf{E}Y \leq \mathbf{E}X$ by Exercise 6.4. By the definition of Y, $\mathbf{E}Y = t\mathbf{P}(X \geq t)$. That is,

$$t\mathbf{P}(X \geq t) \leq \mathbf{E}(X).$$

□

Remark 7.2. A nearly identical proof shows that $\mathbf{P}(X > t) \leq \frac{\mathbf{E}X}{t}$, for all $t > 0$.

Markov's inequality is commonly applied in the following ways.

Corollary 7.3. *Let X be a random variable. Then*

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

More generally, if n is a positive integer, then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

Proof. The first assertion follows immediately by applying Proposition 7.1 to $|X|$. For the second assertion, we use the first assertion to get

$$\mathbf{P}(|X| \geq t) = \mathbf{P}(|X|^n \geq t^n) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

□

The second inequality of Corollary 7.3 is fairly useful, since if many moments of $|X|$ are bounded, then $\mathbf{P}(|X| \geq t)$ decays very rapidly.

Replacing X by $X - \mu$ and taking $n = 2$ in Corollary 7.3 gives:

Corollary 7.4 (**Chebyshev Inequality**). *Let X be a random variable with mean μ. Then*

$$\mathbf{P}(|X - \mu| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}, \qquad \forall\, t > 0.$$

Or, replacing t by $t\sqrt{\mathrm{var}(X)}$,

$$\mathbf{P}(|X - \mu| \geq t\sqrt{\mathrm{var}(X)}) \leq \frac{1}{t^2}, \qquad \forall\, t > 0.$$

Exercise 7.5. Let X be a standard Gaussian random variable. Let $t > 0$ and let n be a positive even integer. Show that

$$\mathbf{P}(X > t) \leq \frac{(n-1)(n-3)\cdots(3)(1)}{t^n}.$$

That is, the function $t \mapsto \mathbf{P}(X > t)$ decays faster than any monomial.

Exercise 7.6. Let X be a random variable. Let $t > 0$. Show that

$$\mathbf{P}(|X| > t) \leq \frac{\mathbf{E}X^4}{t^4}.$$

Exercise 7.7 (The Chernoff Bound). Let X be a random variable and let $r > 0$. Show that, for any $t > 0$,

$$\mathbf{P}(X > r) \leq e^{-tr} M_X(t).$$

Consequently, if $X_1, \ldots, X_n$ are independent random variables with the same CDF, and if $r, t > 0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > r\right) \leq e^{-trn}(M_{X_1}(t))^n.$$

For example, if $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter $0 < p < 1$, and if $r, t > 0$,

$$\mathbf{P}\left(\frac{X_1 + \cdots + X_n}{n} - p > r\right) \leq e^{-trn}(e^{-tp}[pe^t + (1-p)])^n.$$

And if we choose t appropriately, then the quantity $\mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - p)\right| > r\right)$ becomes exponentially small as either n or r become large. That is, $\frac{1}{n}\sum_{i=1}^{n} X_i$ becomes very close to its mean. Importantly, the Chernoff bound is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > r\right) \leq \frac{2p(1-p)}{r}, \quad \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > r\right) \leq \frac{p(1-p)}{nr^2}.$$

Proposition 7.8 (Borel-Cantelli Lemma). *Let $A_1, A_2, \ldots$ be events with $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. Let $B := \{\sum_{n=1}^{\infty} 1_{A_n} = \infty\}$, so that B is the event that infinitely many of the events $A_1, A_2, \ldots$ occur. Then $\mathbf{P}(B) = 0$.*

Proof. For any $n \geq 1$, let 1_{A_n} be a random variable which is 1 if A_n occurs, and 0 otherwise. That is, $1_{A_n}(\omega) = 1$ if $\omega \in A_n$, and $1_{A_n}(\omega) = 0$ if $\omega \notin A_n$. Then $\mathbf{E}(\sum_{n=1}^{\infty} 1_{A_n}) = \sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. So, by Markov's inequality, Proposition 7.1,

$$\mathbf{P}\left(\sum_{n=1}^{\infty} 1_{A_n} \geq t\right) \leq \frac{\sum_{n=1}^{\infty} \mathbf{P}(A_n)}{t}, \qquad \forall\, t > 0.$$

Letting $t \to \infty$ and using Continuity of the Probability Law, Proposition 6.6,

$$\mathbf{P}(B) = \mathbf{P}\left(\sum_{n=1}^{\infty} 1_{A_n} = \infty\right) = \lim_{t\to\infty} \mathbf{P}\left(\sum_{n=1}^{\infty} 1_{A_n} \geq t\right) = 0.$$

□

7.2. **Weak Law of Large Numbers.**

Definition 7.9. Let $X_1, X_2, \ldots$ be random variables. We say that $X_1, X_2, \ldots$ are **identically distributed** if $X_1, X_2, \ldots$ all have the same CDF. That is, $\mathbf{P}(X_i \leq t) = \mathbf{P}(X_j \leq t)$ for all $t \in \mathbb{R}$ and for all positive integers i, j.

Remark 7.10. If $X_1, X_2, \ldots$ are identically distributed random variables, then $\mathbf{E}X_i = \mathbf{E}X_j$ for all positive integers i, j.

We know intuitively that, if the results of independent experiments are averaged, then the average will become close to the expected value of a single experiment. Indeed, one way to intuitively think about expected value is as the average of many repeated experiments. The Law of Large Numbers makes the previous statement rigorous. For now, we only prove a weak version of this statement, though a stronger version will be proven later.

Theorem 7.11 (**Weak Law of Large Numbers**). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables. Assume that $\mu \in \mathbb{R}$ and $\mathbf{E}X_1 = \mu$. Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \varepsilon \right) = 0.$$

Proof. We make the additional assumption that $\mathrm{var}(X_1) < \infty$. Removing this assumption relies on things outside of this class. From Corollary 6.28,

$$\mathrm{var}\left(\frac{X_1 + \cdots + X_n}{n} \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{1}{n} \mathrm{var}(X_1).$$

So, Chebyshev's inequality implies that

$$\mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \varepsilon \right) \leq \frac{1}{n} \varepsilon^{-2} \mathrm{var}(X_1).$$

Letting $n \to \infty$ concludes the proof. □

Remark 7.12. We saw in Exercise 7.7 that the Chernoff bound implies the Weak Law of Large Numbers. However, the Chernoff bound requires the moment generating function to exist and be close to 1 for small $t > 0$, which is a much stronger assumption than what we assumed in Theorem 7.11.

Example 7.13. Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with parameter $1/2$. Let $n := 10^4$, $\varepsilon := 10^{-2}$. Then

$$\mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \frac{1}{2} \right| \geq \frac{1}{100} \right) \leq 10^{-4} 10^4 (1/4) = \frac{1}{4}.$$

7.3. **Convergence in Probability.**

Definition 7.14. We say that a sequence of random variables $Y_1, Y_2, \ldots$ **converges in probability** to a random variable Y if: for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \varepsilon) = 0.$$

More formally, if Ω is the sample space, then $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(\omega \in \Omega\colon\ |Y_n(\omega) - Y(\omega)| > \varepsilon) = 0$.

Remark 7.15. So, the Weak Law of Large numbers says: if X_1, X_2 are independent identically distributed random variables with $\mu := \mathbf{E}X_1 \in \mathbb{R}$, then the random variables $\frac{X_1+\cdots+X_n}{n}$ converge in probability to the constant μ.

Example 7.16. For any $n \geq 1$, let Y_n be a random variable such that $\mathbf{P}(Y_n = n^2) = 1/n$, and $\mathbf{P}(Y_n = 0) = 1 - 1/n$. Then $Y_1, Y_2, \ldots$ converges in probability to 0. For any $\varepsilon > 0$,

$$\mathbf{P}(|Y_n - 0| > \varepsilon) = \mathbf{P}(|Y_n| > \varepsilon) = \mathbf{P}(Y_n = n^2) = 1/n.$$

Therefore, $\lim_{n\to\infty} \mathbf{P}(|Y_n - 0| > \varepsilon) = 0$.

However, note that convergence in probability does not imply convergence in expected value, since $\lim_{n\to\infty} \mathbf{E}Y_n = \lim_{n\to\infty} n = \infty$, whereas the expected value of 0 is just 0.

Proposition 7.17 (**Uniqueness of the Limit**). *Suppose $Y_1, Y_2, \ldots$ converges in probability to Y. Also, suppose $Y_1, Y_2, \ldots$ converges in probability to Z. Then $\mathbf{P}(Z \neq Y) = 0$.*

Proof. From the triangle inequality, for any $n \geq 1$,

$$|Z - Y| = |Z - Y_n + Y_n - Y| \leq |Z - Y_n| + |Y - Y_n|.$$

So, for any $\varepsilon > 0$, if $|Z - Y| \geq \varepsilon$, then either $|Z - Y_n| \geq \varepsilon/2$ or $|Y - Y_n| \geq \varepsilon/2$. That is, for any $\varepsilon > 0$ and for any $n \geq 1$,

$$\begin{aligned}&\{\omega \in \Omega\colon\ |Z(\omega) - Y(\omega)| \geq \varepsilon\}\\ &\quad\subseteq \{\omega \in \Omega\colon\ |Z(\omega) - Y_n(\omega)| \geq \varepsilon/2\} \cup \{\omega \in \Omega\colon\ |Y(\omega) - Y_n(\omega)| \geq \varepsilon/2\}.\end{aligned}$$

Therefore, for any $\varepsilon > 0$ and for any $n \geq 1$,

$$\mathbf{P}(|Z - Y| \geq \varepsilon) \leq \mathbf{P}(|Z - Y_n| \geq \varepsilon/2) + \mathbf{P}(|Y - Y_n| \geq \varepsilon/2).$$

The left side does not depend on n. So, letting $n \to \infty$, we get $\mathbf{P}(|Z - Y| \geq \varepsilon) = 0$, for all $\varepsilon > 0$. Now,

$$\{Z \neq Y\} \subseteq \cup_{t=1}^{\infty} \{|Z - Y| \geq 1/t\}.$$

Therefore, $\mathbf{P}(Z \neq Y) \leq \sum_{t=1}^{\infty} \mathbf{P}(|Z - Y| \geq 1/t) = 0$. So, $\mathbf{P}(Z \neq Y) = 0$. □

Exercise 7.18. Let $X_1, X_2, \ldots$ be independent random variables, each with exponential distribution with parameter $\lambda = 1$. For any $n \geq 1$, let $Y_n := \max(X_1, \ldots, X_n)$. Let $0 < a < 1 < b$. Show that $\mathbf{P}(Y_n \leq a \log n) \to 0$ as $n \to \infty$, and $\mathbf{P}(Y_n \leq b \log n) \to 1$ as $n \to \infty$. Conclude that $Y_n / \log n$ converges to 1 in probability as $n \to \infty$.

Exercise 7.19. We say that random variables $X_1, X_2, \ldots$ converge to a random variable X in L_2 if

$$\lim_{n\to\infty} \mathbf{E}\,|X_n - X|^2 = 0.$$

Show that, if $X_1, X_2, \ldots$ converge to X in L_2, then $X_1, X_2, \ldots$ converges to X in probability.

Is the converse true? Prove your assertion.

Exercise 7.20. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables such that $\mathbf{E}\,|X_1| < \infty$ and $\operatorname{var}(X_1) < \infty$. For any $n \geq 1$, define

$$Y_n := \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Show that $Y_1, Y_2, \ldots$ converges in probability. Express the limit in terms of $\mathbf{E}X_1$ and $\operatorname{var}(X_1)$.

7.4. **Central Limit Theorem.** The following is a stronger version of Theorem 6.39.

Theorem 7.21 (Lévy Continuity Theorem). *Let $X_1, X_2, \ldots$ be random variables and let X be a random variable. For any fixed $t \in \mathbb{R}$, assume that $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$. Assume also that $\phi_X(t)$ is continuous at $t = 0$. Then for any fixed $t \in \mathbb{R}$ such that $\mathbf{P}(X \le t)$ is continuous, we have $\lim_{n\to\infty} \mathbf{P}(X_n \le t) = \mathbf{P}(X \le t)$.*

In particular, if X, Y are random variables with $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$, and if $\phi_X(t)$ is continuous at $t = 0$, then X, Y are identically distributed.

We are finally able to prove the generalization of the De Moivre Laplace Theorem, Theorem 6.1, to arbitrary random variables.

Theorem 7.22 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables. Let Z be a standard Gaussian random variable. Let $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. Assume that $\mathbf{E}X_1 = \mu$ and $\operatorname{var}(X_1) = \sigma^2$. Then for any $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le t\right) = \int_{-\infty}^{t} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \mathbf{P}(Z \le t).$$

Remark 7.23. The random variable $\frac{X_1+\cdots+X_n-\mu n}{\sigma\sqrt{n}}$ has mean zero and variance 1, just like the standard Gaussian Z.

Exercise 7.24. Let $f, g, h \colon \mathbb{R} \to \mathbb{R}$. We use the notation $f(t) = o(g(t))\ \forall\ t \in \mathbb{R}$ to denote $\lim_{t\to 0} \left|\frac{f(t)}{g(t)}\right| = 0$. For example, if $f(t) = t^3\ \forall\ t \in \mathbb{R}$, then $f(t) = o(t^2)$, since $\lim_{t\to 0} |\frac{f(t)}{t^2}| = \lim_{t\to 0} |t| = 0$. Show: (i) if $f(t) = o(g(t))$ and if $h(t) = o(g(t))$, then $(f + h)(t) = o(g(t))$. (ii) If c is any nonzero constant, then $o(cg(t)) = o(g(t))$. (iii) $\lim_{t\to 0} g(t)o(1/g(t)) = 0$. (iv) $\lim_{t\to 0} o(g(t))/g(t) = 0$. (v) $o(g(t) + o(g(t))) = o(g(t))$.

Proof. For any $j \ge 1$, let $Y_j := (X_j - \mu)/\sigma$. Then $Y_1, Y_2, \ldots$ are independent and identically distributed, $\mathbf{E}Y_j = 0$ and $\mathbf{E}Y_j^2 = 1$, $\forall\ j \ge 1$. We will show that $\lim_{n\to\infty} \mathbf{P}(\frac{Y_1+\cdots+Y_n}{\sqrt{n}} \le t) = \mathbf{P}(Z \le t)$, $\forall\ t \in \mathbb{R}$. From Theorem 7.21 and Proposition 6.49, it suffices to show:

$$\lim_{n\to\infty} \mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \mathbf{E}e^{itZ} = e^{-t^2/2}, \qquad \forall\, t \in \mathbb{R}.$$

From Proposition 6.48,

$$\mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \prod_{j=1}^{n} \mathbf{E}e^{itY_j/\sqrt{n}} = (\mathbf{E}e^{itY_1/\sqrt{n}})^n.$$

We make the additional assumption that $\mathbf{E}\,|X_1|^3 < \infty$, so that $\mathbf{E}\,|Y_1|^3 < \infty$ and we can apply Exercise 6.47. (As remarked in Exercise 6.47, this assumption is not needed for the conclusion of Exercise 6.47 to hold.) By Exercise 6.47, and using $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^2 = 1$,

$$\mathbf{E}e^{itY_1/\sqrt{n}} = 1 + \frac{it}{\sqrt{n}}\mathbf{E}Y_1 - \frac{t^2}{2n}\mathbf{E}Y_1^2 + o(t^2/n) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Therefore,

$$\mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Taking logarithms, using $\log(1+x) = x + o(x)$ for $-1 < x < 1$, and using Exercise 7.24,

$$\log \mathbf{E} e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = n \log\left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right) = -\frac{t^2}{2} + n \cdot o\left(\frac{t^2}{n}\right).$$

Letting $n \to \infty$ and using Exercise 7.24(iii) completes the proof. □

Definition 7.25 (**Convergence in Distribution**)**.** Let $X, X_1, X_2, \ldots$ be random variables. We say that $X_1, X_2, \ldots$ **converge in distribution** to X if, for any $t \in \mathbb{R}$ such that the CDF of X is continuous at t,

$$\lim_{n\to\infty} \mathbf{P}(X_n \leq t) = \mathbf{P}(X \leq t).$$

So, the Central Limit Theorem, Theorem 7.22, says: if $X_1, X_2, \ldots$ are independent, identically distributed random variables with $\mu := \mathbf{E}X_1$ and $\sigma^2 := \mathrm{Var}(X_1)$ with $\sigma > 0$, then the random variables $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ converge in distribution to the standard Gaussian random variable. This fact is rather remarkable, since it holds no matter what distribution X_1 has! In this sense, the Gaussian random variable is "universal."

Exercise 7.26. This exercise demonstrates that geometry in high dimensions is different than geometry in low dimensions.

Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Let $\|x\| := \sqrt{x_1^2 + \cdots + x_n^2}$. Let $\varepsilon > 0$. Show that for all sufficiently large n, "most" of the cube $[-1,1]^n$ is contained in the annulus

$$A := \{x \in \mathbb{R}^n \colon (1-\varepsilon)\sqrt{n/3} \leq \|x\| \leq (1+\varepsilon)\sqrt{n/3}\}.$$

That is, if $X_1, \ldots, X_n$ are each independent and identically distributed in $[-1,1]$, then for n sufficiently large

$$\mathbf{P}((X_1, \ldots, X_n) \in A) \geq 1 - \varepsilon.$$

(Hint: apply the weak law of large numbers to $X_1^2, \ldots, X_n^2$.)

Exercise 7.27 (**Confidence Intervals**)**.** Among 625 members of a bank chosen uniformly at random among all bank members, it was found that 25 had a savings account. Give an interval of the form $[a,b]$ where $0 \leq a, b \leq 625$ are integers, such that with about 95% certainty, if we sample 625 bank members independently and uniformly at random (from a very large bank membership), then the number of these people with savings accounts lies in the interval $[a,b]$. (Hint: if Y is a standard Gaussian random variable, then $\mathbf{P}(-2 \leq Y \leq 2) \approx .95$.)

Exercise 7.28 (**Hypothesis Testing**)**.** Suppose we run a casino, and we want to test whether or not a particular roulette wheel is biased. Let p be the probability that red results from one spin of the roulette wheel. Using statistical terminology, "$p = 18/38$" is the null hypothesis, and "$p \neq 18/38$" is the alternative hypothesis. (On a standard roulette wheel, 18 of the 38 spaces are red.) For any $i \geq 1$, let $X_i = 1$ if the i^{th} spin is red, and let $X_i = 0$ otherwise.

Let $\mu := \mathbf{E}X_1$ and let $\sigma := \sqrt{\mathrm{var}(X_1)}$. If the null hypothesis is true, and if Y is a standard Gaussian random variable

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right| \geq 2\right) = \mathbf{P}(|Y| \geq 2) \approx .05.$$

To test the null hypothesis, we spin the wheel n times. In our test, we reject the null hypothesis if $|X_1 + \cdots + X_n - n\mu| > 2\sigma\sqrt{n}$. Rejecting the null hypothesis when it is true is

called a type I error. In this test, we set the type I error percentage to be 5%. (The type I error percentage is closely related to the p-value.)

Suppose we spin the wheel $n = 3800$ times and we get red 1868 times. Is the wheel biased? That is, can we reject the null hypothesis with around 95% certainty?

7.5. **Strong Law of Large Numbers.** We recall Example 7.16. For any $n \geq 1$, let Y_n be a random variable such that $\mathbf{P}(Y_n = n^2) = 1/n$, and $\mathbf{P}(Y_n = 0) = 1 - 1/n$. In Example 7.16, we showed that $Y_1, Y_2, \ldots$ converges in probability to $X := 0$. In fact, these random variables also converge in distribution to 0. Let $t \in \mathbb{R}$. Then by the definition of $Y_1, Y_2, \ldots$,

$$\lim_{n\to\infty} \mathbf{P}(X_n \leq t) = \begin{cases} \lim_{n\to\infty}(1-1/n) & \text{, if } t \geq 0 \\ 0 & \text{, if } t < 0 \end{cases} = \begin{cases} 1 & \text{, if } t \geq 0 \\ 0 & \text{, if } t < 0 \end{cases} = \mathbf{P}(X \leq t).$$

In fact, convergence in probability always implies convergence in distribution, but the converse is false.

Exercise 7.29. Suppose random variables $X_1, X_2, \ldots$ converge in probability to a random variable X. Prove that $X_1, X_2, \ldots$ converge in distribution to X.

Then, show that the converse is false.

By Exercise 7.29, we see that the convergence guaranteed by the Central Limit Theorem is weaker than convergence in probability. We might hope to upgrade the Central Limit Theorem to get the stronger convergence in probability, but unfortunately this is impossible.

Exercise 7.30. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$. For any $n \geq 1$, define

$$S_n := \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

The Central Limit Theorem says that S_n converges in distribution to a standard Gaussian random variable. We show that S_n does not converge in probability to any random variable. The intuition here is that if S_n did converge in probability to a random variable Z, then when n is large, S_n is close to Z, $Y_n := \frac{\sqrt{2}S_{2n} - S_n}{\sqrt{2}-1}$ is close to Z, but S_n and Y_n are independent. And this cannot happen.

Proceed as follows. Assume that S_n converges in probability to Z.

- Let $\varepsilon > 0$. For n very large (depending on ε), we have $\mathbf{P}(|S_n - Z| > \varepsilon) < \varepsilon$ and $\mathbf{P}(|Y_n - Z| > \varepsilon) < \varepsilon$.
- Show that $\mathbf{P}(S_n > 0, Y_n > 0)$ is around $1/4$, using independence and the Central Limit Theorem.
- From the first item, show $\mathbf{P}(S_n > 0 | Z > \varepsilon) > 1 - \varepsilon$, $\mathbf{P}(Y_n > 0 | Z > \varepsilon) > 1 - \varepsilon$, so $\mathbf{P}(S_n > 0, Y_n > 0 | Z > \varepsilon) > 1 - 2\varepsilon$.
- Without loss of generality, for ε small, we have $\mathbf{P}(Z > \varepsilon) > 4/9$.
- By conditioning on $Z > \varepsilon$, show that $\mathbf{P}(S_n > 0, Y_n > 0)$ is at least $3/8$, when n is large.

The Weak Law of Large Numbers, Theorem 7.11, showed that the average $\frac{X_1 + \cdots + X_n}{n}$ of independent identically distributed random variables with finite mean converges to the mean in probability. We can upgrade this convergence in probability to a stronger notion of convergence, which we now define.

Definition 7.31 (**Almost Sure Convergence**). We say that random variables $X_1, X_2, \ldots$ converge **almost surely** (or **with probability one**) to a random variable X if

$$\mathbf{P}(\lim_{n\to\infty} X_n = X) = 1.$$

More rigorously, if Ω is the sample space, then $\mathbf{P}(\{\omega \in \Omega\colon\ \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$

Exercise 7.32. Let $X_1, X_2, \ldots$ be random variables that converge almost surely to a random variable X. That is,

$$\mathbf{P}(\lim_{n\to\infty} X_n = X) = 1.$$

Show that $X_1, X_2, \ldots$ converges in probability to X in the following way.

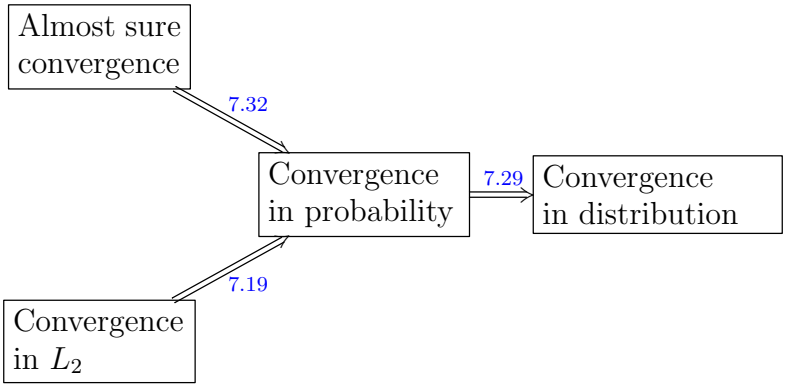
- For any $\varepsilon > 0$ and for any positive integer n, let

$$A_{n,\varepsilon} := \bigcup_{m=n}^{\infty} \{\omega \in \Omega\colon\ |X_m(\omega) - X(\omega)| > \varepsilon\}.$$

  Show that $A_{n,\varepsilon} \supseteq A_{n+1,\varepsilon} \supseteq A_{n+2,\varepsilon} \supseteq \cdots$.
- Show that $\mathbf{P}(\cap_{n=1}^{\infty} A_{n,\varepsilon}) = 0$.
- Using Continuity of the Probability Law, deduce that $\lim_{n\to\infty} \mathbf{P}(A_{n,\varepsilon}) = 0$.

Now, show that the converse is false. That is, find random variables $X_1, X_2, \ldots$ that converge in probability to X, but where $X_1, X_2, \ldots$ do not converge to X almost surely.

Remark 7.33. The following table summarizes our different notions of convergence of random variables. That is, the following table summarizes the implications of Exercises 7.19, 7.29 and 7.32.



Remark 7.34. Almost sure convergence does not imply convergence in L_2, and convergence in L_2 does not imply almost sure convergence.

To see the first, assertion, recall the random variables $Y_1, Y_2, \ldots$ constructed in Example 7.16. Then $Y_1, Y_2, \ldots$ converges almost surely to 0, since $\lim_{n\to\infty} Y_n(t) = 0$ for all $t \in (0, 1]$, so $\mathbf{P}(\lim_{n\to\infty} Y_n = 0) = \mathbf{P}((0, 1]) = 1$. On the other hand, $Y_1, Y_2, \ldots$ does not converge in L_2 to 0, since $\mathbf{E}\,|Y_n - 0|^2 = \mathbf{E}Y_n^2 = n^4/n = n^3$, so $\lim_{n\to\infty} \mathbf{E}\,|Y_n - 0|^2 \neq 0$.

We now show that convergence in L_2 does not imply almost sure convergence. Let $\mathbf{P}$ be the uniform probability law on $[1, 2]$. For any positive integer n, define $X_n\colon [1, 2] \to \mathbb{R}$ as follows. Let $j = j(n)$ be the nonnegative integer such that $2^j \leq n < 2^{j+1}$. Let $X_n(t) := 1$ if $t \in [n2^{-j}, (n+1)2^{-j}]$, and let $X_n(t) := 0$ otherwise. We claim that $X_1, X_2, \ldots$ converges to 0 in L_2, but $X_1, X_2, \ldots$ does not converge almost surely to 0. Note that $\mathbf{E}\,|X_n - 0|^2 = \mathbf{E}X_n^2 = 2^{-j}$, and as $n \to \infty$, $j \to \infty$, so that $\lim_{n\to\infty} \mathbf{E}\,|X_n - 0|^2 = 0$. However, for any $t \in [0, 1]$,

there exist infinitely many values of n such that $X_n(t) = 1$ and infinitely many values of n such that $X_n(t) = 0$. Therefore, $\lim_{n\to\infty} X_n(t)$ does not exist, for every $t \in [0,1]$. That is, $X_1, X_2, \ldots$ does not converge almost surely to any random variable.

Exercise 7.35. Using the Central Limit Theorem, prove the Weak Law of Large Numbers.

Theorem 7.36 (Strong Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume that $\mu = \mathbf{E}X_1$. Then*

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Proof. We prove the Theorem under the stronger assumption that $\mathbf{E}X_1^4 < \infty$. For any $j \geq 1$, let $Y_j := X_j - \mu$. We are required to show $\mathbf{P}\left(\lim_{n\to\infty} \frac{Y_1+\cdots+Y_n}{n} = 0\right) = 1$. Note that $Y_1, Y_2, \ldots$ are independent identically distributed random variables with $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^4 < \infty$. We compute

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 = \sum_{1\leq i,j,k,\ell\leq n} \mathbf{E}Y_iY_jY_kY_\ell.$$

By independence, terms with $i \neq j = k = \ell$ vanish, since they become $\mathbf{E}Y_iY_jY_kY_\ell = \mathbf{E}Y_i\mathbf{E}Y_j^3 = 0$. Terms with i, j, k, ℓ distinct also vanish, since $\mathbf{E}Y_iY_jY_kY_\ell = \mathbf{E}Y_i\mathbf{E}Y_j\mathbf{E}Y_k\mathbf{E}Y_\ell = 0$. The remaining nonvanishing terms are $i = j = k = \ell$ and the six permutations of $i = j \neq k = \ell$. That is,

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 = n\mathbf{E}Y_1^4 + 6[n(n-1)/2](\mathbf{E}Y_1^2)^2.$$

By Jensen's Inequality, Exercise 4.23,

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 \leq n\mathbf{E}Y_1^4 + 3n(n-1)\mathbf{E}Y_1^4 \leq 4n^2\mathbf{E}Y_1^4. \qquad (*)$$

By Markov's Inequality, Proposition 7.1, for any $t > 0$,

$$\mathbf{P}\left(\left|\frac{Y_1 + \cdots + Y_n}{n}\right| > t\right) \leq \frac{\mathbf{E}(Y_1 + \cdots + Y_n)^4}{t^4n^4} \overset{(*)}{\leq} \frac{4\mathbf{E}Y_1^4}{t^4n^2}.$$

So $\sum_{n=1}^{\infty} \mathbf{P}(\left|\frac{Y_1+\cdots+Y_n}{n}\right| > t) < \infty$ and by Borel-Cantelli, Proposition 7.8, $\forall\ t > 0$,

$$\mathbf{P}\left(\left|\frac{Y_1 + \cdots + Y_n}{n}\right| > t \text{ for infinitely many } n \geq 1\right) = 0.$$

Since this holds for any $t > 0$, we conclude that $\frac{Y_1+\cdots+Y_n}{n}$ converges almost surely to 0. □

Remark 7.37. The Strong Law of Large Numbers implies the Weak Law of Large Numbers by Exercise 7.32.

Exercise 7.38 (Renewal Theory). Let $t_1, t_2, \ldots$ be positive, independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume $\mathbf{E}t_1 = \mu$. For any positive integer j, we interpret t_j as the lifetime of the j^{th} lightbulb (before burning out, at which point it is replaced by the $(j+1)^{st}$ lightbulb). For any $n \geq 1$, let $T_n := t_1 + \cdots + t_n$ be the total lifetime of the first n lightbulbs. For any positive integer t, let $N_t := \min\{n \geq 1 \colon T_n \geq t\}$ be the number of lightbulbs that have been used up until time t. Show that N_t/t converges almost surely to $1/\mu$ as $t \to \infty$. (Hint: if c, t are positive integers, then $\{N_t \leq ct\} = \{T_{ct} \geq t\}$. Apply the Strong Law to T_{ct}.)

Exercise 7.39 (**Playing Monopoly Forever**). Let $t_1, t_2, \ldots$ be independent random variables, all of which are uniform on $\{1, 2, 3, 4, 5, 6\}$. For any positive integer j, we think of t_j as the result of rolling a single fair six-sided die. For any $n \geq 1$, let $T_n = t_1 + \cdots + t_n$ be the total number of spaces that have been moved after the n^{th} roll. (We think of each roll as the amount of moves forward of a game piece on a very large Monopoly game board.) For any positive integer t, let $N_t := \min\{n \geq 1 : T_n \geq t\}$ be the number of rolls needed to get t spaces away from the start. Using Exercise 7.38, show that N_t/t converges almost surely to $2/7$ as $t \to \infty$.

Exercise 7.40 (**Random Numbers are Normal**). Let X be a uniformly distributed random variable on $(0, 1)$. Let X_1 be the first digit in the decimal expansion of X. Let X_2 be the second digit in the decimal expansion of X. And so on.

- Show that the random variables $X_1, X_2, \ldots$ are uniform on $\{0, 1, 2, \ldots, 9\}$ and independent.
- Fix $m \in \{0, 1, 2, \ldots, 9\}$. Using the Strong Law of Large Numbers, show that with probability one, the fraction of appearances of the number m in the first n digits of X converges to $1/10$ as $n \to \infty$.

(Optional): Show that for any ordered finite set of digits of length k, the fraction of appearances of this set of digits in the first n digits of X converges to 10^{-k} as $n \to \infty$. (You already proved the case $k = 1$ above.) That is, a randomly chosen number in $(0, 1)$ is normal. On the other hand, if we just pick some number such that $\sqrt{2} - 1$, then it may not be easy to say whether or not that number is normal.

(As an optional exercise, try to explicitly write down a normal number. This may not be so easy to do, even though a random number in $(0, 1)$ satisfies this property!)

Exercise 7.41. Let $X_1, X_2, \ldots$ be random variables with mean zero and variance one. The Strong Law of Large Numbers says that $\frac{1}{n}(X_1 + \cdots + X_n)$ converges almost surely to zero. The Central Limit Theorem says that $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ converges in distribution to a standard Gaussian random variable. But what happens if we divide by some other power of n? This Exercise gives a partial answer to this question.

Let $\varepsilon > 0$. Show that

$$\frac{X_1 + \cdots + X_n}{n^{1/2}(\log n)^{(1/2)+\varepsilon}}$$

converges to zero almost surely as $n \to \infty$. (Hint: Re-do the proof of the Strong Law of Large Numbers, but divide by $n^{1/2}(\log n)^{(1/2)+\varepsilon}$ instead of n.)

8. Appendix: Notation

Let n, m be a positive integers. Let A, B be sets contained in a universal set Ω.

$\mathbb{R}$ denotes the set of real numbers

$\in$ means "is an element of." For example, $2 \in \mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\forall$ means "for all"

$\exists$ means "there exists"

$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \colon x_i \in \mathbb{R} \, \forall \, 1 \leq i \leq n\}$

$f \colon A \to B$ means f is a function with domain A and range B. For example,

$f \colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall \, a \in A$, we have $a \in B$, so A is contained in B

$A \setminus B := \{a \in A \colon a \notin B\}$

$A^c := \Omega \setminus A$, the complement of A in Ω

$A \cap B$ denotes the intersection of A and B

$A \cup B$ denotes the union of A and B

$\mathbf{P}$ denotes a probability law on Ω

$\mathbf{P}(A|B)$ denotes the conditional probability of A, given B.

Let $a_1, \ldots, a_n$ be real numbers. Let n be a positive integer.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_{n-1} + a_n.$$

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdots a_{n-1} \cdot a_n.$$

$\min(a_1, a_2)$ denotes the minimum of a_1 and a_2.

$\max(a_1, a_2)$ denotes the maximum of a_1 and a_2.

Let X be a discrete random variable on a sample space Ω, so that $X \colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x \in \mathbb{R}$. Let $A \subseteq \Omega$. Let Y be another discrete random variable

$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{\omega \in \Omega \colon X(\omega) = x\}), \, \forall \, x \in \mathbb{R}$

the Probability Mass Function (PMF) of X

$\mathbf{E}(X)$ denotes the expected value of X

$\operatorname{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2$, the variance of X

$\sigma_X = \sqrt{\operatorname{var}(X)}$, the standard deviation of X

$X|A$ denotes the random variable X conditioned on the event A.

$\mathbf{E}(X|A)$ denotes the expected value of X conditioned on the event A.

$1_A\colon \Omega \to \{0,1\}$, denotes the indicator function of A, so that

$$1_A(\omega) = \begin{cases} 1 & \text{, if } \omega \in A \\ 0 & \text{, otherwise.} \end{cases}$$

Let X, Y be a continuous random variables on a sample space Ω, so that $X, Y\colon \Omega \to \mathbb{R}$. Let $-\infty \le a \le b \le \infty$, $-\infty \le c \le d \le \infty$. Let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$.

$f_X\colon \mathbb{R} \to [0,\infty)$ denotes the Probability Density Function (PDF) of X, so

$$\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)dx$$

$f_{X,Y}\colon \mathbb{R} \to [0,\infty)$ denotes the joint PDF of X and Y, so

$$\mathbf{P}(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

$f_{X|A}$ denotes the Conditional PDF of X given A

$\mathbf{E}(X|A)$ denotes the expected value of X conditioned on the event A.

Let X be a random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x, t \in \mathbb{R}$. Let $i := \sqrt{-1}$.

$F_X(x) = \mathbf{P}(X \le x) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) \le x\})$
the Cumulative Distibution Function (CDF) of X.

$M_X(t) = \mathbf{E}e^{tX}$ denotes the Moment Generating Function of X at $t \in \mathbb{R}$

$\phi_X(t) = \mathbf{E}e^{itX}$ denotes the Characteristic Function (or Fourier Transform) of X at $t \in \mathbb{R}$

Let $g, h\colon \mathbb{Z} \to \mathbb{R}$. Let $t \in \mathbb{Z}$.

$$(g * h)(t) = \sum_{j \in \mathbb{Z}} g(j)h(t-j) \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{Z}$$

Let $g, h\colon \mathbb{R} \to \mathbb{R}$. Let $t \in \mathbb{R}$.

$$(g * h)(t) = \int_{-\infty}^{\infty} g(x)h(t-x)dx \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{R}$$

Let $f, g\colon \mathbb{R} \to \mathbb{R}$. We use the notation $f(t) = o(g(t))$, $\forall\, t \in \mathbb{R}$ to denote $\lim_{t\to 0} \left|\frac{f(t)}{g(t)}\right| = 0$.

USC Mathematics, Los Angeles, CA
E-mail address: stevenmheilman@gmail.com