MA 581: Introduction to Probability

Complete Lecture Notes

Work in Progress

Instructor Wancheng Lin

Summer 2026

Acknowledgments. These course notes were developed together with our fantastic students — Yang Lu, Eris Pil, and Madi Kim — whose lecture notes and feedback shaped this document. Special thanks also to Professor Solesne Bourguin and Professor Konstantinos Spiliopoulos for their advice and for sharing materials.

Contents

1 Mathematical Preliminaries: Sets and Cardinality

Probability theory is built on the language of *sets*. Before introducing any probability, we collect the necessary definitions here so that the notes are self-contained.

1.1 What Is a Set?

Definition 1.1 (Set)**.** A **set** is any well-defined collection of distinct objects, called its **elements** (or **members**). We write $x \in A$ to mean "x is an element of A", and $x \notin A$ for the negation.

Sets are typically denoted by capital letters $A, B, C, \Omega, \ldots$ and described either by *roster notation* (listing all elements) or *set-builder notation*:

$$A = \{1, 3, 5\}, \qquad B = \{x \in \mathbb{R} : x \geq 0\}.$$

Definition 1.2 (Empty Set)**.** The **empty set** $\emptyset$ (also written $\{\}$) is the unique set containing *no* elements whatsoever. For every object x, we have $x \notin \emptyset$.

1.2 Set Relations

Definition 1.3 (Subset and Equality)**.** We say A is a **subset** of B, written $A \subseteq B$, if every element of A is also an element of B:

$$A \subseteq B \iff \forall x,\ x \in A \implies x \in B.$$

Two sets are **equal**, $A = B$, if and only if $A \subseteq B$ and $B \subseteq A$.

Remark 1.1. The empty set $\emptyset$ is a subset of *every* set: $\emptyset \subseteq A$ for all A. This is vacuously true — there are no elements in $\emptyset$ that fail to belong to A.

1.3 Set Operations

Throughout, we work inside a fixed **universal set** Ω (in probability, the sample space).

Definition 1.4 (Union, Intersection, Complement, Difference)**.** Let A and B be subsets of Ω.

- **Union:** $A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}$.
- **Intersection:** $A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\}$.
- **Complement:** $A^C = \Omega \setminus A = \{x \in \Omega : x \notin A\}$.
- **Set difference:** $A \setminus B = A \cap B^C = \{x \in \Omega : x \in A \text{ and } x \notin B\}$.

We say A and B are **disjoint** (mutually exclusive) if $A \cap B = \emptyset$.

These operations extend to countable (and uncountable) families: for a collection $\{A_i\}_{i \in I}$,

$$\bigcup_{i \in I} A_i = \{x : x \in A_i \text{ for some } i\}, \qquad \bigcap_{i \in I} A_i = \{x : x \in A_i \text{ for all } i\}.$$

Proposition 1.1 (Key Set Identities). *The following identities hold for all subsets $A, B, C \subseteq \Omega$:*

(i) ***De Morgan's Laws:*** $(A \cup B)^C = A^C \cap B^C, \quad (A \cap B)^C = A^C \cup B^C.$

(ii) ***Distributive Laws:*** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$

(iii) ***Partition identity:*** $A = (A \cap B) \cup (A \cap B^C)$ *(the two pieces are disjoint).*

1.4 Cardinality: How Many Elements Does a Set Have?

The **cardinality** of a set A, written $|A|$ or $\#A$, is the number of elements it contains. For finite sets this is simply a counting number: $|\{a, b, c\}| = 3$, $|\emptyset| = 0$. For infinite sets the picture is more subtle, and understanding it is *essential* for probability.

1.4.1 Finite Sets: Cardinality as a Counting Number

If a set has n elements for some $n \in \mathbb{N}_0$, we call it **finite** with $|A| = n$.

Example 1.1. $|\{H, T\}| = 2; \quad |\{1, 2, 3, 4, 5, 6\}| = 6; \quad |\emptyset| = 0.$

1.4.2 The Question That Opens Infinity: What If the List Never Ends?

Now suppose we try to count the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$. We start: $1, 2, 3, \dots$ — but the list never terminates. So $|\mathbb{N}|$ is not any finite number; it is *infinite*. At this point a natural — and deep — question arises:

Are all infinite sets "the same size"?

The answer, discovered by Cantor in the 19th century, is **no**. There are genuinely different sizes of infinity, and probability theory cares deeply about which kind a sample space has. The right notion of "same size" for sets (finite or infinite) is a **bijection**.

Definition 1.5 (Bijection and Equal Cardinality). Two sets A and B have **equal cardinality**, $|A| = |B|$, if there exists a **bijection** $f : A \to B$ — a function that is both injective (one-to-one) and surjective (onto). Intuitively, f pairs every element of A with a unique element of B, leaving nothing out on either side.

1.4.3 Two Flavours of Infinity

Countably infinite sets. The "smallest" infinite cardinality is that of $\mathbb{N}$. A set is **countably infinite** if it can be put in bijection with $\mathbb{N}$ — equivalently, its elements can be arranged in an infinite list $a_1, a_2, a_3, \dots$ that eventually reaches every element.

Example 1.2 (The Even Numbers Are Countably Infinite). The map $n \mapsto 2n$ is a bijection $\mathbb{N} \to \{2, 4, 6, \dots\}$, so both sets have the same cardinality despite $\{2, 4, 6, \dots\}$ looking "half as big".

Example 1.3 (The Integers $\mathbb{Z}$ Are Countably Infinite)**.** List them as $0, 1, -1, 2, -2, 3, -3, \ldots$ — this exhaustive sequence is the required bijection with $\mathbb{N}$.

Example 1.4 (The Rationals $\mathbb{Q}$ Are Countably Infinite)**.** This is more surprising: despite $\mathbb{Q}$ being *dense* in $\mathbb{R}$ (between any two rationals there is another rational), the rationals can still be systematically listed. The standard proof uses Cantor's diagonal enumeration of fractions p/q. We will not reproduce it here; see Wikipedia: Countable set for a complete account.

Uncountably infinite sets. Some infinite sets are so vast that *no* infinite list can exhaust them.

Example 1.5 (The Reals $\mathbb{R}$ Are Uncountably Infinite)**.** Cantor's diagonal argument shows that $[0, 1]$ cannot be put in bijection with $\mathbb{N}$: any purported list $r_1, r_2, r_3, \ldots$ misses at least one element. See Wikipedia: Cantor's diagonal argument. We call $\mathbb{R}$ (and any interval $[a, b]$) **uncountably infinite**.

Definition 1.6 (Countable vs. Uncountable)**.** A set is **countable** if it is either finite or countably infinite. A set is **uncountable** if it cannot be put in bijection with $\mathbb{N}$.

The cardinality hierarchy relevant to us is:

$$\text{finite} \subset \text{countably infinite} \subset \text{uncountably infinite}.$$

Set	Cardinality type	Listable?
$\{1, 2, 3, 4, 5, 6\}$	Finite	Yes (6 elements)
$\mathbb{N} = \{1, 2, 3, \ldots\}$	Countably infinite	Yes (natural order)
$\mathbb{Z}$	Countably infinite	Yes (interleaving)
$\mathbb{Q}$	Countably infinite	Yes (Cantor diagonal)
$\mathbb{R}$, or any interval $[a, b]$	Uncountably infinite	No

Remark 1.2 (Why This Matters for Probability)**.** Kolmogorov's third axiom (Section 2.6) says that probability is *countably additive*: we can sum probabilities over a countable collection of disjoint events. Attempting to "sum" over an *uncountable* index set — such as all real numbers in $[0, 1]$ — is *not permitted* by the axioms. This distinction will be critical when we ask whether individual points can have positive probability in continuous sample spaces (Section 4.1).

2 Lecture 1: June 29, 2026 — Sample Spaces & Kolmogorov's Axioms

Experiments can broadly be classified into two distinct categories: **deterministic** and **random**.

2.1 Deterministic vs. Random Experiments

Definition 2.1 (Deterministic Experiment)**.** An experiment whose outcomes are predictable and certain given the initial conditions; the outcomes are completely governed by physical laws.

Definition 2.2 (Random Experiment)**.** An experiment where individual outcomes cannot be predicted with absolute certainty beforehand, but the set of all possible outcomes is known, and these outcomes possess associated long-term relative frequencies or probabilities.

Example 2.1. The most classical examples:

- $\mathcal{E}_1 = \{\text{Rolling a standard six-sided die}\}; \quad \Omega_1 = \{1,2,3,4,5,6\}.$
- $\mathcal{E}_2 = \{\text{Flipping a fair coin}\}; \quad \Omega_2 = \{H,T\}.$

2.2 Elements of a Random Experiment

To mathematically formalize a random experiment, we must define the following structural elements:

1. **Specify the experiment** $\mathcal{E}$.
2. **Specify the sample space** Ω: the set of *all* possible *elementary outcomes* of $\mathcal{E}$.
3. **Define events**: subsets $A \subseteq \Omega$.
4. **Assign probabilities** to each defined event (governed by Kolmogorov's Axioms, Section 2.6).

2.3 Further Examples of Sample Spaces

- $\mathcal{E}_1 = \{\text{Toss a coin}\}, \quad \Omega_1 = \{H,T\}$
- $\mathcal{E}_2 = \{\text{Toss 2 coins}\}, \quad \Omega_2 = \{(H,H),(H,T),(T,H),(T,T)\}$
- $\mathcal{E}_3 = \{\text{Lifetime of a lightbulb (hours)}\}, \quad \Omega_3 = [0,\infty)$
- $\mathcal{E}_4 = \{\text{Number of particles emitted}\}, \quad \Omega_4 = \{0,1,2,\dots\} = \mathbb{N}_0$
- $\mathcal{E}_5 = \{\text{France vs. Norway: final score}\}, \quad \Omega_5 = \{(f,n) \mid f,n \in \mathbb{N}_0\}$

2.4 Events Language vs. Set-Theoretic Language

Event language	**Set language**
Something happens (certainty)	Ω (sample space)
Nothing happens (impossibility)	$\emptyset$ (empty set)
Event A occurs	A
A implies B	$A \subseteq B$
A does *not* occur	$A^C = \Omega \setminus A$
A *or* B occurs	$A \cup B$
A *and* B occur	$A \cap B$
A and B are mutually exclusive	$A \cap B = \emptyset$
Pairwise mutually exclusive collection	$A_i \cap A_j = \emptyset$ for all $i \neq j$

2.5 Fundamental Set-Theoretic Laws

- **De Morgan's Laws:**

$$(A \cup B)^C = A^C \cap B^C, \qquad (A \cap B)^C = A^C \cup B^C.$$

- **Distributive Laws:**

$$A \cap \left(\bigcup_i B_i\right) = \bigcup_i (A \cap B_i), \qquad A \cup \left(\bigcap_i B_i\right) = \bigcap_i (A \cup B_i).$$

- **Partition of an Event (foundation for total probability):** For any events A and B,

$$A = A \cap \Omega = A \cap (B \cup B^C) = (A \cap B) \cup (A \cap B^C),$$

and the two pieces are disjoint (Proposition 1.1(iii)).

> **Remark 2.1.** This technique is powerful when computing $\mathbb{P}(A)$ directly is hard but conditioning on a supplementary event B (or a full partition $\{B_i\}$) is easier. For instance, to find the probability France wins (event A), decompose A over all specific winning scorelines $B_i = \{(f,n) : f - n = i,\ f > n\}$; each B_i is more tractable individually.

As a concrete illustration, Ω_5 partitions into three exhaustive disjoint outcomes:

- $A_{\text{win}} = \{(f,n) \in \Omega_5 \mid f > n\}$ (France wins)
- $B_{\text{win}} = \{(f,n) \in \Omega_5 \mid f < n\}$ (Norway wins)
- $C_{\text{draw}} = \{(f,n) \in \Omega_5 \mid f = n\}$ (Draw)

These three sets are pairwise disjoint and their union is all of Ω_5.

2.6 Kolmogorov's Axioms of Probability

Goal: Construct a function $\mathbb{P}$ assigning a probability to each event $A \subseteq \Omega$, in a way that is logically consistent.

Probability Measure (Kolmogorov's Axioms)

A function $\mathbb{P}$ is a **probability measure** on Ω if it satisfies:

1. **Non-negativity:** $\mathbb{P}(A) \geq 0$ for every event A.
2. **Normalization:** $\mathbb{P}(\Omega) = 1$.
3. **Countable Additivity:** For any *countable* collection of pairwise disjoint events $A_1, A_2, \ldots$ (with $A_i \cap A_j = \emptyset$ for $i \neq j$),

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Remark 2.2 (Finite Additivity is Included in Axiom 3)**.** Since any finite collection is countable (Section 1.4), Axiom 3 directly covers the finite case: for pairwise disjoint $A_1, \ldots, A_N$,

$$\mathbb{P}\left(\bigcup_{i=1}^{N} A_i\right) = \sum_{i=1}^{N} \mathbb{P}(A_i).$$

In practice this is the form used most often — whenever we write $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for disjoint A, B, we are invoking Axiom 3 with $N = 2$. The restriction to *countable* (rather than uncountable) families is deliberate and load-bearing; we will see exactly why it matters in Section 4.1.

3 Lecture 2: June 30, 2026 — Consequences of the Axioms

3.1 Consequences of the Kolmogorov Axioms

Proposition 3.1. *The three axioms imply:*

1. ***Complement Rule:*** $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$.
2. ***Empty Set:*** $\mathbb{P}(\emptyset) = 0$.
3. ***Monotonicity:*** $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$.
4. ***Addition Rule:*** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
5. ***Sub-additivity (Boole's Inequality):*** $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Proof. We prove each statement in order, citing only the axioms and previously established results.

(1) Complement Rule: $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$**.**

By definition of complement, A and A^C are disjoint and $A \cup A^C = \Omega$. Since we are taking the union of *two* sets — a finite, hence countable, collection — Axiom 3 applies:

$$\mathbb{P}(\Omega) = \mathbb{P}(A \cup A^C) = \mathbb{P}(A) + \mathbb{P}(A^C).$$

By Axiom 2, $\mathbb{P}(\Omega) = 1$, so $\mathbb{P}(A) + \mathbb{P}(A^C) = 1$, and rearranging gives $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$.

(2) Empty Set: $\mathbb{P}(\emptyset) = 0$**.**

We use two elementary set facts:
- A and $\emptyset$ are disjoint: $A \cap \emptyset = \emptyset$.
- Adjoining the empty set changes nothing: $A \cup \emptyset = A$.

Since A and $\emptyset$ form a finite (hence countable) disjoint collection, Axiom 3 gives:

$$\mathbb{P}(A) = \mathbb{P}(A \cup \emptyset) = \mathbb{P}(A) + \mathbb{P}(\emptyset).$$

Cancelling $\mathbb{P}(A)$ from both sides: $\mathbb{P}(\emptyset) = 0$.

(3) Monotonicity: $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$**.**

Assume $A \subseteq B$. Every element of B either belongs to A or to $B \setminus A$, giving the disjoint decomposition:

$$B = A \cup (B \cap A^C), \qquad A \cap (B \cap A^C) = \emptyset.$$

Applying Axiom 3 and then Axiom 1:

$$\mathbb{P}(B) = \mathbb{P}(A) + \underbrace{\mathbb{P}(B \cap A^C)}_{\geq 0} \geq \mathbb{P}(A).$$

(4) Addition Rule: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$**.**

Observe the two disjoint decompositions:

$$A \cup B = A \cup (B \cap A^C), \qquad B = (B \cap A) \cup (B \cap A^C).$$

Applying Axiom 3 to each:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^C), \tag{1}$$

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^C). \tag{2}$$

Subtracting (2) from (1) eliminates $\mathbb{P}(B \cap A^C)$:

$$\mathbb{P}(A \cup B) - \mathbb{P}(B) = \mathbb{P}(A) - \mathbb{P}(A \cap B),$$

and rearranging gives $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

(5) Sub-additivity (Boole's Inequality): $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.
From (4): $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. By Axiom 1, $\mathbb{P}(A \cap B) \geq 0$, so:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \underbrace{\mathbb{P}(A \cap B)}_{\geq 0} \leq \mathbb{P}(A) + \mathbb{P}(B).$$

By induction this extends to any finite collection, and to countably infinite collections by a limit argument:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i). \qquad \blacksquare$$

3.2 Inclusion–Exclusion Principle

The addition rule for two events already reveals the pattern: summing $\mathbb{P}(A) + \mathbb{P}(B)$ over-counts outcomes in $A \cap B$, so we subtract the overlap once. For three events the same logic cascades:

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = {} & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ & + \mathbb{P}(A \cap B \cap C).\end{aligned}$$

Theorem 3.2 (Inclusion–Exclusion Principle). *For events $A_1, A_2, \ldots, A_n$,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{m=1}^{n} (-1)^{m+1} \sum_{1 \leq k_1 < k_2 < \cdots < k_m \leq n} \mathbb{P}(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m}).$$

Expanding level by level: sum all singletons, subtract all pairwise intersections, add all triple intersections, and so on with alternating signs.

Remark 3.1 (Why the Signs Alternate: Over-count then Correct). Any outcome ω belonging to exactly m of the sets A_i is counted $\binom{m}{1}$ times in the first sum, removed $\binom{m}{2}$ times in the second, restored $\binom{m}{3}$ times in the third, and so on. The binomial identity $\sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k} = 1$ guarantees the net count is exactly 1 for every such ω.

Fundamental mode of reasoning. Inclusion–exclusion is a *sieve*: when $\mathbb{P}(\bigcup A_i)$ is hard to compute directly because the sets overlap in complicated ways, decompose it into intersections, which are typically easier to evaluate individually. Truncating after the first term recovers Boole's inequality $\mathbb{P}(\bigcup_i A_i) \leq \sum_i \mathbb{P}(A_i)$ as a free upper bound.

Example 3.1 (The Job-Hunt Sieve). You are scanning a job market. Three desirable traits and their estimated probabilities are:

Event	Trait	Probability
A	Great salary	0.50
B	Remote work	0.40
C	Healthy culture	0.30
$A \cap B$	Salary *and* remote	0.20
$A \cap C$	Salary *and* good culture	0.10
$B \cap C$	Remote *and* good culture	0.15
$A \cap B \cap C$	All three (the "unicorn")	0.05

Naively summing $0.5 + 0.4 + 0.3 = 1.2 > 1$ is impossible — the overlaps are double-counted. Applying inclusion–exclusion:

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) &= (0.5 + 0.4 + 0.3) - (0.2 + 0.1 + 0.15) + 0.05 \\ &= 1.2 - 0.45 + 0.05 \;=\; \mathbf{0.80}.\end{aligned}$$

There is an 80% chance a random offer has *at least one* redeeming trait; equivalently, 20% of jobs are simultaneously low-paying, in-office, and toxic — the region where no circle intersects anything worthwhile.

4 Lecture 3: Geometric Probability & Law of Total Probability

4.1 Why Singleton Events Must Have Probability Zero in Continuous Spaces

A Billion-Dollar Thought Experiment

Imagine you are offered the following bet at a casino.

> **The Billion-Dollar Wheel.** A dealer spins a perfectly continuous, frictionless wheel whose face is the interval $[0, 1]$. The wheel stops at a position ω chosen *uniformly at random* — no point on $[0, 1]$ is favoured over any other. *Before* the wheel is spun, you must name an exact number $\omega^* \in [0, 1]$. If the wheel lands *exactly* on ω^*, you win \$1,000,000,000.

Intuition. The sample space $\Omega = [0, 1]$ is an uncountable continuum; you are betting on one specific element $\{\omega^*\}$ out of uncountably many indistinguishable points. Intuitively your chance is vanishingly small — in fact, *exactly zero*. The theorem below makes this rigorous. We will return immediately after the proof to a second, more vivid version of this bet — a dart thrown at a disc — which will open the door to *geometric probability* in Section 4.3.

Theorem 4.1 (Singletons Have Zero Probability Under the Uniform Law)**.** *For the continuous uniform probability measure on $\Omega = [0, 1]$, every singleton event has probability zero:*

$$\mathbb{P}(\{\omega\}) = 0 \qquad \text{for all } \omega \in [0, 1].$$

Remark 4.1 (Why the Naive Argument Fails — and What to Do Instead)**.** One might try: "if every point has probability $\varepsilon > 0$, summing over all of $[0, 1]$ exceeds 1, a contradiction." This sounds right but silently invokes Axiom 3 over the *uncountable* index set $[0, 1]$. Axiom 3 applies *only to countable collections*; the sum

$$\mathbb{P}(\Omega) \overset{?}{=} \sum_{\omega \in [0,1]} \mathbb{P}(\{\omega\})$$

is not defined within our framework. The correct move is to *restrict to a countable subset*: the rationals $S := \mathbb{Q} \cap [0, 1]$, which are countably infinite (Section 1.4), so Axiom 3 applies legitimately there.

Proof. Suppose for contradiction $\mathbb{P}(\{\omega_0\}) = \varepsilon > 0$ for some $\omega_0 \in [0, 1]$. Since the wheel is *uniform*, symmetry forces

$$\mathbb{P}(\{\omega\}) = \varepsilon \qquad \text{for all } \omega \in [0, 1].$$

Let $S := \mathbb{Q} \cap [0, 1] = \{q_1, q_2, q_3, \dots\}$, countably infinite by Section 1.4. The singletons $\{q_i\}$ are pairwise disjoint and their index set is countable, so **Axiom 3 applies**:

$$\mathbb{P}(S) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} \{q_i\}\right) = \sum_{i=1}^{\infty} \mathbb{P}(\{q_i\}) = \sum_{i=1}^{\infty} \varepsilon = \infty.$$

But $S \subseteq \Omega$, so Monotonicity gives $\mathbb{P}(S) \leq \mathbb{P}(\Omega) = 1$, yielding $\infty \leq 1$, a contradiction. Hence $\mathbb{P}(\{\omega\}) = 0$ for every $\omega \in [0, 1]$. ■

Remark 4.2 (Probability Zero $\neq$ Impossible)**.** $\mathbb{P}(\{\omega^*\}) = 0$ does *not* mean the wheel cannot land on ω^*. After one spin it *will* land somewhere, and that exact point has probability zero. In continuous spaces, **probability zero** and **impossibility** are genuinely different concepts. Probability is not carried by points but by *intervals*:

$$\mathbb{P}([a,b]) = b - a, \qquad 0 \le a < b \le 1,$$

and since $\mathbb{P}(\{a\}) = \mathbb{P}(\{b\}) = 0$, endpoints are irrelevant: $\mathbb{P}([a,b]) = \mathbb{P}((a,b)) = \mathbb{P}([a,b)) = \mathbb{P}((a,b])$.

The house can safely offer a billion-dollar prize for hitting ω^* exactly because $\mathbb{P}(\{\omega^*\}) = 0$ — it will never be paid out. To have a *non-zero* chance you must bet on an *interval*, say $[0.4, 0.6]$, which carries probability 0.2. And if the wheel is replaced by a dart thrown uniformly at a disc, the same logic applies — but now probability is proportional to *area*, not length. That observation is the starting point of Section 4.3.

4.2 From the Dart Board to Geometric Probability

Recall the dart variant of our billion-dollar bet. A professional player throws a dart at a unit disc; the landing point (x, y) is chosen *uniformly* over the disc. We now know $\mathbb{P}(\{(x,y)\}) = 0$ for every candidate point. But notice what changes the moment we ask a slightly different question:

What is the probability the dart lands within distance r of the centre?

Now we are asking about a *region* — a disc of radius r — not a single point. The answer is immediate once we accept one natural principle: on a uniform sample space, *probability is proportional to area*. This is the idea behind **geometric probability**.

4.3 Geometric Probability (Continuous Uniform Distributions)

Definition 4.1 (Geometric Probability Model)**.** Let $\Omega \subseteq \mathbb{R}^n$ be a region with finite positive measure $|\Omega|$ (length/area/volume). For any event $E \subseteq \Omega$,

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|},$$

where $|\cdot|$ denotes length ($n = 1$), area ($n = 2$), or volume ($n = 3$).

Remark 4.3. All three Kolmogorov axioms are satisfied: non-negativity is immediate; $\mathbb{P}(\Omega) = |\Omega|/|\Omega| = 1$; countable additivity follows from the additivity of Lebesgue measure over disjoint regions. The singleton result of Section 4.1 is subsumed here: a single point has zero length/area/volume, hence zero probability.

Remark 4.4 (The Unit Interval Revisited)**.** Setting $\Omega = [0,1]$ with $|\Omega| = 1$ recovers the standard continuous uniform distribution. Singletons have measure zero (Section 4.1), and probability is carried by intervals: $\mathbb{P}((a,b]) = b - a$. More general measurable sets inherit their probability via the theory of measure.

4.3.1 Worked Example: Dart on a Unit Disc

Example 4.1 (Billion-Dollar Dart: Safe Zone). A dart is thrown uniformly at a unit disc. Let E be the event that the dart lands *more than* $\frac{1}{2}$ *unit from the centre*. Find $\mathbb{P}(E)$.

Solution. The sample space is the closed unit disc $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ with area $|\Omega| = \pi$. The favourable region is the annulus outside the inner disc of radius $\frac{1}{2}$:

$$E = \left\{(x, y) \in \Omega : x^2 + y^2 > \tfrac{1}{4}\right\}, \qquad |E| = \pi(1)^2 - \pi\left(\tfrac{1}{2}\right)^2 = \pi - \tfrac{\pi}{4} = \tfrac{3\pi}{4}.$$

Therefore $\mathbb{P}(E) = \dfrac{|E|}{|\Omega|} = \dfrac{3\pi/4}{\pi} = \dfrac{3}{4}$. Three-quarters of the disc lies outside the inner half-radius region; probability is completely determined by area. ■

4.4 Partitions and the Law of Total Probability

Definition 4.2 (Partition of Ω). A countable collection $\{A_1, A_2, \dots\}$ is a **partition** of Ω if

(i) **Pairwise disjoint:** $A_i \cap A_j = \emptyset$ for all $i \neq j$.

(ii) **Exhaustive:** $\bigcup_i A_i = \Omega$.

Proposition 4.2 (Law of Total Probability). *Let $\{A_1, A_2, \dots\}$ be a partition of Ω and B any event. Then*

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i \cap B).$$

Proof. Since $\{A_i\}$ partitions Ω: $B = B \cap \Omega = \bigcup_i (A_i \cap B)$. The sets $A_i \cap B$ are pairwise disjoint and the collection is countable, so Axiom 3 gives

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcup_i (A_i \cap B)\right) = \sum_i \mathbb{P}(A_i \cap B).$$ ■

Remark 4.5. With conditional probabilities, $\mathbb{P}(A_i \cap B) = \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)$, giving the classical form $\mathbb{P}(B) = \sum_i \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)$, to be developed when we introduce conditional probability.

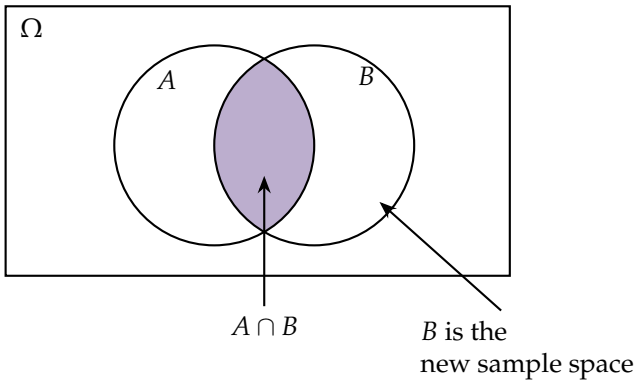
5 Lecture 5: July 6, 2026 — Conditional Probability, Total Probability, and Bayes' Rule

5.1 Conditional Probability

Definition 5.1 (Conditional Probability). For events A, B with $\mathbb{P}(B) > 0$, the **conditional probability of A given B** is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\text{intersection}}{\text{conditioning set}}.$$

Conditioning on B means we have learned that B occurred: the sample space contracts from Ω down to B, and we renormalise so that the new total probability is again 1.



5.2 The Multiplication Rule and Its Extension

Rearranging the definition gives the **multiplication rule**:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A).$$

Iterating it chains through any number of events.

Theorem 5.1 (General Multiplication (Chain) Rule). *For events A, B, C (with the conditioning events having positive probability),*

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}((A \cap B) \cap C) = \mathbb{P}(C \mid A \cap B)\,\mathbb{P}(B \mid A)\,\mathbb{P}(A).$$

In general,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\ \mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Example 5.1 (Five Clubs, Two Ways). Draw 5 cards from a 52-card deck without replacement. Let $E = \{\text{all five cards are clubs } \clubsuit\}$. Compute $\mathbb{P}(E)$ two ways.

Solution. **Method 1 (counting, equally likely outcomes).** All $\binom{52}{5}$ hands are equally likely; the

favourable hands choose 5 of the 13 clubs:

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{\binom{13}{5}}{\binom{52}{5}} = \frac{1287}{2{,}598{,}960} \approx 4.95 \times 10^{-4}.$$

Method 2 (multiplication rule, card by card). Let $A_i = \{\text{the } i\text{-th card drawn is a club}\}$. Then $E = A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$, and the chain rule gives

$$\mathbb{P}(E) = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} \approx 4.95 \times 10^{-4}.$$

Each factor conditions on the previous draws: once k clubs are gone, $13 - k$ clubs remain among $52 - k$ cards. The two methods agree. ■

> **Remark 5.1** (Two Related but Different Questions)**.** "All five cards are clubs" (≈ 0.0005, above) is not the same question as "the fifth card is a club." The latter is a single-position event: by symmetry every position in a shuffled deck is equally likely to hold any card, so $\mathbb{P}(\text{5th card is a club}) = \frac{13}{52} = \frac{1}{4}$, exactly as for the first card.

5.3 The Law of Total Probability

> **Theorem 5.2** (Law of Total Probability)**.** *If $B_1, B_2, \ldots, B_n$ form a **partition** of Ω (pairwise disjoint, union $= \Omega$, each $\mathbb{P}(B_i) > 0$), then for any event A,*
>
> $$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)\,.$$
>
> *Thus $\mathbb{P}(A)$ is the* weighted average *of the conditional probabilities $\mathbb{P}(A \mid B_i)$, weighted by how likely each scenario B_i is.*

> **Remark 5.2.** This is the probabilistic form of the partition identity from Lecture 1: writing $A = \bigcup_i (A \cap B_i)$ as a disjoint union and applying Axiom 3 gives $\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i)$; the multiplication rule then rewrites each term as $\mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)$.

> **Example 5.2** (Is the Second Student Male?)**.** A class has 40 students: 17 men and 23 women. Two students are picked one after another, uniformly at random, *without replacement*. Let $E = \{\text{the second student is male}\}$. Find $\mathbb{P}(E)$.

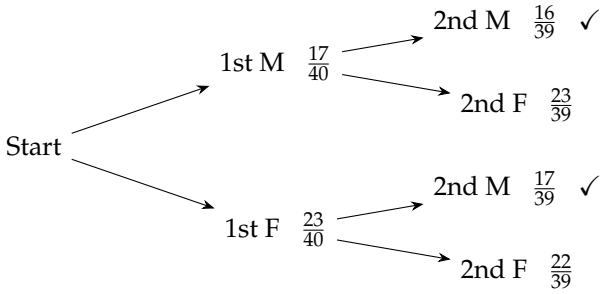
Solution. Partition by the first pick: $B_1 = \{\text{1st is male}\}$, $B_2 = \{\text{1st is female}\}$. These are disjoint and exhaustive, with $\mathbb{P}(B_1) = \frac{17}{40}$, $\mathbb{P}(B_2) = \frac{23}{40}$. Conditional on the first pick, 39 students remain:

$$\mathbb{P}(E \mid B_1) = \frac{16}{39}, \qquad \mathbb{P}(E \mid B_2) = \frac{17}{39}.$$

By the Law of Total Probability,

$$\mathbb{P}(E) = \frac{16}{39} \cdot \frac{17}{40} + \frac{17}{39} \cdot \frac{23}{40} = \frac{17\,(16 + 23)}{39 \cdot 40} = \frac{17 \cdot 39}{39 \cdot 40} = \frac{17}{40} = 0.425.$$

Notice the answer equals $\mathbb{P}(B_1) = \frac{17}{40}$: by symmetry the second draw is just as likely to be male as the first.



■

5.4 Bayes' Rule

Theorem 5.3 (Bayes' Rule). *For events A, B with $\mathbb{P}(A), \mathbb{P}(B) > 0$,*

$$\underbrace{\mathbb{P}(A \mid B)}_{\text{posterior (updated belief)}} = \frac{\overbrace{\mathbb{P}(B \mid A)}^{\text{multiplication rule}} \overbrace{\mathbb{P}(A)}^{\text{prior}}}{\underbrace{\mathbb{P}(B)}_{\text{evidence}}}.$$

In words: update the prior belief $\mathbb{P}(A)$ using the new evidence B to obtain the posterior belief $\mathbb{P}(A \mid B)$.

Remark 5.3 (Where Bayes' Rule Comes From: Two Ways of Cutting Up $A \cap B$). Bayes' rule is not an extra axiom — it falls straight out of the *definition* of conditional probability, written two ways.

(1) By definition, $\mathbb{P}(A \mid B) := \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, so *as an extension*, $\mathbb{P}(A \mid B)\ \mathbb{P}(B) = \mathbb{P}(A \cap B)$.

(2) Equally, $\mathbb{P}(B \mid A) := \dfrac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$, so *as an extension*, $\mathbb{P}(B \mid A)\ \mathbb{P}(A) = \mathbb{P}(B \cap A)$.

(3) Since $A \cap B = B \cap A$, the right-hand sides of (1) and (2) agree:

$$\mathbb{P}(A \mid B)\ \mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\ \mathbb{P}(A)\,.$$

Dividing through by $\mathbb{P}(B)$ gives exactly Bayes' rule. In other words, $\mathbb{P}(A \cap B)$ is a single number that can be *built* two ways — go through B first, or go through A first — and equating those two constructions is the whole derivation.

The updating story. Call A the event under analysis and $\mathbb{P}(A)$ our **prior**: what we believe about A *before* any new information arrives. Now a new event B occurs — fresh data, a test result, a diagnosis, a season change — and we ask how this should revise our belief in A. The revised belief is the **posterior**, $\mathbb{P}(A \mid B)$. For example, let $A = \{\text{it rains tomorrow}\}$: on its own $\mathbb{P}(A)$ might be small, but if $B = \{\text{it is currently rainy season}\}$, then learning B should push our belief in A upward, and Bayes' rule tells us exactly by how much. Nothing about the world changes when we condition — what changes is *our information*, and Bayes' rule is

the precise bookkeeping for that change:

$$\underbrace{\mathbb{P}(A)}_{\text{before } B} \xrightarrow{\text{observe } B} \underbrace{\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}}_{\text{after } B}.$$

The factor $\mathbb{P}(B \mid A)\,/\mathbb{P}(B)$ is the "surprise ratio": if B is more likely under A than it is on average, this ratio exceeds 1 and the posterior is pulled *above* the prior; if B is less likely under A than on average, the posterior is pulled below.

Example 5.3 (Lung Cancer and Smoking). Define $L = \{\text{this person has lung cancer}\}$ and $S = \{\text{this person is a smoker}\}$, with sample space $\Omega = L \cup L^C$. Suppose we know

$$\mathbb{P}(L) = 0.07, \qquad \mathbb{P}(S \mid L) = 0.9, \qquad \mathbb{P}\left(S \mid L^C\right) = 0.253.$$

Here $\mathbb{P}(L)$ is our **prior** belief that a random person has lung cancer, before we know anything about their smoking habits. We now learn the new event S — this person smokes — and ask how that updates our belief in L: what is $\mathbb{P}(L \mid S)$?

Solution. By Bayes' rule, together with the Law of Total Probability to expand the denominator $\mathbb{P}(S) = \mathbb{P}(S \mid L)\,\mathbb{P}(L) + \mathbb{P}\left(S \mid L^C\right)\mathbb{P}\left(L^C\right)$,

$$\mathbb{P}(L \mid S) = \frac{\mathbb{P}(S \mid L)\;\mathbb{P}(L)}{\mathbb{P}(S)} = \frac{\mathbb{P}(S \mid L)\;\mathbb{P}(L)}{\mathbb{P}(S \mid L)\;\mathbb{P}(L) + \mathbb{P}(S \mid L^C)\;\mathbb{P}(L^C)} = \frac{(0.9)(0.07)}{(0.9)(0.07) + (0.253)(0.93)} \approx 0.211.$$

The prior probability of lung cancer was only 7%. Observing that the person smokes — new evidence S — pulls the posterior up to about 21%, roughly a threefold increase. This is exactly the updating story above: smoking is much more common among people with lung cancer (90%) than in the general population ($\mathbb{P}(S) \approx 25.8\%$), so the "surprise ratio" $\mathbb{P}(S \mid L)\,/\mathbb{P}(S)$ is greater than 1 and the evidence pulls belief in L upward. ■

Theorem 5.4 (Generalised Bayes' Rule (Two or More Causes)). *Let $A_1, \ldots, A_n$ form a partition of Ω (the competing "causes"), and let B be observed evidence with $\mathbb{P}(B) > 0$. Then for each i,*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B \mid A_j)\,\mathbb{P}\left(A_j\right)}.$$

The denominator is precisely the Law of Total Probability applied to $\mathbb{P}(B)$.

Remark 5.4. Typical use: the causes are competing diagnoses, e.g. $A_1 = \text{flu}$, $A_2 = \text{COVID}$, $A_3 = \text{other}$, and the evidence is $B = \text{fever}$. Then $\mathbb{P}(\text{flu} \mid \text{fever})$ and $\mathbb{P}(\text{COVID} \mid \text{fever})$ are computed from the same denominator $\mathbb{P}(\text{fever})$, and the posteriors over all causes sum to 1.

6 Lecture 6: July 7, 2026 — Independence and Random Variables

6.1 Independence

Definition 6.1 (Independence). Events A and B (with positive probability) are **independent**, written $A \perp B$, when the occurrence of A does not change the probability of B:

$$\mathbb{P}(B \mid A) = \mathbb{P}(B) \qquad \Longleftrightarrow \qquad \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)\,.$$

The product form is the standard definition: it is symmetric in A, B and remains meaningful even when $\mathbb{P}(A) = 0$.

Example 6.1 (Two Dice). Roll two fair dice. Let $A = \{\text{die 1 shows 1}\}$ and $B = \{\text{die 2 shows 2}\}$. Are A and B independent?

Solution. We have $\mathbb{P}(B \mid A) = \frac{1}{6} = \mathbb{P}(B)$: knowing die 1 tells us nothing about die 2. Equivalently, $\mathbb{P}(A \cap B) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbb{P}(A)\,\mathbb{P}(B)$, so $A \perp B$. ■

Remark 6.1 (Independence Is Not Disjointness). Independent events can (and typically do) overlap — one event simply carries no information about the other. Disjointness is a different notion: two disjoint events with positive probability are *never* independent, since knowing A occurred forces $\mathbb{P}(B \mid A) = 0 \neq \mathbb{P}(B)$.

6.2 Random Variables

Motivating example. Toss two fair coins and count the number of heads. This count is not known before the experiment, can only take the values $\{0, 1, 2\}$, and is a *number* computed from the outcome. This is the prototype of a random variable.

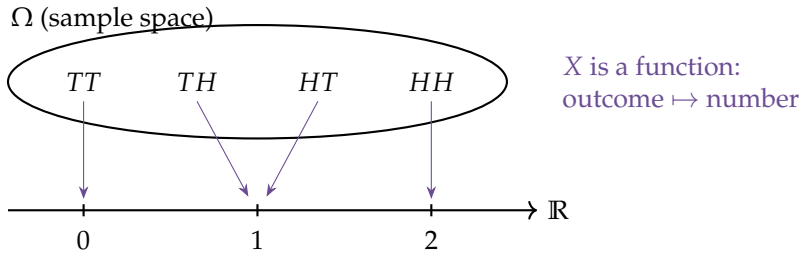
Definition 6.2 (Random Variable). A **random variable** is a function from the sample space to the real numbers,

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

assigning a number $X(\omega)$ to each outcome ω.

For the two-coin toss with $X =$ number of heads,

$$X(HH) = 2, \qquad X(HT) = 1, \qquad X(TH) = 1, \qquad X(TT) = 0.$$



Definition 6.3 (Probability Mass Function (PMF)). For a discrete random variable X, the **probability mass function** is

$$p_X(x) = \mathbb{P}(X = x),$$

the probability that X takes the value x. It satisfies $p_X(x) \geq 0$ and $\sum_x p_X(x) = 1$, the sum taken over the possible values (the *support*).

For the two-coin example the four outcomes are equally likely, so

$$\mathbb{P}(X = 0) = \tfrac{1}{4}, \qquad \mathbb{P}(X = 1) = \tfrac{2}{4}, \qquad \mathbb{P}(X = 2) = \tfrac{1}{4}, \qquad \mathbb{P}(X \notin \{0, 1, 2\}) = 0.$$

6.3 Types of Discrete Random Variables

Definition 6.4 (Discrete Uniform Distribution). $X \sim \text{Uniform}(n)$: X takes one of n equally likely values, with PMF

$$\mathbb{P}(X = x) = \frac{1}{n} \quad \text{for each of the } n \text{ values.}$$

For instance, roll a fair die and let X be the outcome; then $\mathbb{P}(X = x) = \frac{1}{6}$ for $x \in \{1, \ldots, 6\}$, i.e. $X \sim \text{Uniform}(6)$.

Definition 6.5 (Bernoulli Distribution). $X \sim \text{Bernoulli}(p)$: a single trial with only two possible outcomes (either/or), coded $X : \Omega \to \{0, 1\}$ with

$$\mathbb{P}(X = 1) = \mathbb{P}(\text{success}) = p, \qquad \mathbb{P}(X = 0) = \mathbb{P}(\text{failure}) = 1 - p.$$

Compactly, $\mathbb{P}(X = x) = p^x (1 - p)^{1-x}$ for $x \in \{0, 1\}$.

Definition 6.6 (Binomial Distribution). $X \sim \text{Bin}(n, p)$: the number of successes when a yes/no trial is repeated n times *independently*, each with the same success probability p. Its PMF is

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Here $\binom{n}{x}$ counts *which* trials succeed, while $p^x (1 - p)^{n-x}$ is the probability of each such pattern.

Example 6.2 (Broken Cookies from Warren Towers). Buy 20 cookies from the Warren Towers dining hall — of famously poor quality: each cookie is broken with probability $\mathbb{P}(\text{broken}) = 0.2$, independently. Let X be the number of broken cookies, so $X \sim \text{Bin}(20, 0.2)$.

(a) What is the probability that exactly 4 cookies are broken?
(b) What is the probability that at least 3 are broken?

Solution. Take success = broken ($p = 0.2$), failure = intact (0.8).

(a) Direct plug-in:

$$\mathbb{P}(X = 4) = \binom{20}{4}(0.2)^4(0.8)^{16} = 4845 \times 0.0016 \times 0.02815 \approx 0.218.$$

(b) "At least 3" means $X \in \{3, 4, \ldots, 20\}$ — eighteen terms — so use the complement:

$$\begin{aligned}\mathbb{P}(X \geq 3) &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) \\ &= 1 - \binom{20}{0}(0.2)^0(0.8)^{20} - \binom{20}{1}(0.2)^1(0.8)^{19} - \binom{20}{2}(0.2)^2(0.8)^{18} \\ &\approx 1 - 0.0115 - 0.0576 - 0.1369 = 0.794.\end{aligned}$$

■

7 Lecture 7: July 8, 2026 — Hypergeometric, Geometric, and Negative Binomial

7.1 Binomial = Sampling *With* Replacement

If a population of N items contains M "successes" and we sample n items *with replacement*, the draws are independent Bernoulli trials with

$$p = \frac{M}{N} = \frac{\#\text{successes}}{\#\text{total}},$$

so the number of successes is binomial:

$$X \sim \text{Bin}\left(n, \tfrac{M}{N}\right), \qquad \mathbb{P}(X = x) = \binom{n}{x}\left(\frac{M}{N}\right)^x\left(1 - \frac{M}{N}\right)^{n-x}.$$

7.2 Hypergeometric = Sampling *Without* Replacement

Without replacement the draws are *not* independent — each draw changes what remains — and the count follows a different law.

Definition 7.1 (Hypergeometric Distribution). Sample n items *without replacement* from a population of N items, of which M are successes. Let X be the number of successes in the sample. Then $X \sim \text{Hyper}(N, M, n)$, with PMF

$$\mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

valid for the feasible values $\max(0,\, n - (N - M)) \leq x \leq \min(n, M)$.

Example 7.1 (All Three Inspected Items Are Good). A shipment has $N = 100$ items, of which $M = 95$ are good (good rate $p = 0.95$). Inspect $n = 3$ items drawn without replacement. What is the probability that all three are good ($x = 3$)?

Solution. Take success = good item:

$$\mathbb{P}(X=3)=\frac{\binom{95}{3}\binom{100-95}{3-3}}{\binom{100}{3}}=\frac{\binom{95}{3}\binom{5}{0}}{\binom{100}{3}}=\frac{138{,}415\times 1}{161{,}700}\approx 0.856.$$

As a sanity check, the chain rule gives the same value: $\frac{95}{100}\cdot\frac{94}{99}\cdot\frac{93}{98}\approx 0.856$. ■

Theorem 7.1 (Binomial Approximation to the Hypergeometric). *If the population is much larger than the sample (rule of thumb $n \le 0.05\,N$), removing a few items barely changes the composition, and*

$$\mathbb{P}(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}\xrightarrow[N\to\infty,\ M/N\to p]{}\binom{n}{x}p^x(1-p)^{n-x}.$$

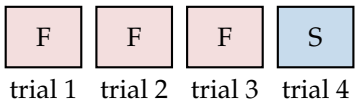
Sampling without replacement from a huge population is approximately sampling with replacement.

7.3 Geometric Distribution: Waiting for the First Success

Definition 7.2 (Geometric Distribution). Repeat independent Bernoulli(p) trials until the *first success*; let X be the trial on which it occurs. Then $X \sim \text{Geom}(p)$ with PMF

$$\mathbb{P}(X=x)=(1-p)^{x-1}\,p, \qquad x=1,2,3,\ldots$$

The formula reads: fail, fail, …, fail ($x-1$ times), then succeed.



$$\mathbb{P}(X=4)=(1-p)^3\,p$$

Example 7.2 (First Broken Cookie). Warren Towers cookies are broken independently with probability $p = 0.2$. What is the probability that the *first* broken cookie is the third one inspected?

Solution. Here $X \sim \text{Geom}(0.2)$, so

$$\mathbb{P}(X=3)=(0.8)^2(0.2)=0.128.$$

■

7.4 Negative Binomial Distribution: Waiting for the r-th Success

Motivating question. When do I get the third broken cookie? This extends the geometric distribution from the first success to the r-th.

Definition 7.3 (Negative Binomial Distribution). Repeat independent Bernoulli(p) trials until the *r-th success*; let X be the trial on which it occurs. Then $X \sim \text{NegBin}(r,p)$ with PMF

$$\mathbb{P}(X=x)=\binom{x-1}{r-1}p^r(1-p)^{x-r}, \qquad x=r,\,r+1,\,r+2,\,\ldots$$

In particular, $\text{Geom}(p) = \text{NegBin}(1, p)$.

Remark 7.1 (Where the PMF Comes From). For the r-th success to land exactly on trial x: trial x itself must be a success (factor p), and among the first $x-1$ trials there must be exactly $r-1$ successes in any positions — $\binom{x-1}{r-1}$ patterns, each with probability $p^{r-1}(1-p)^{x-r}$. Multiplying gives $\binom{x-1}{r-1}p^r(1-p)^{x-r}$. The support starts at $x = r$, since the r-th success cannot appear before trial r.

Example 7.3 (Third Broken Cookie). With $p = 0.2$: what is the probability that the *third* broken cookie is the tenth cookie inspected?

Solution. Here $X \sim \text{NegBin}(3, 0.2)$ and $x = 10$:

$$\mathbb{P}(X = 10) = \binom{9}{2}(0.2)^3(0.8)^7 = 36 \times 0.008 \times 0.2097 \approx 0.0604.$$

■

8 Lecture 8: July 9, 2026 — The Poisson Distribution, Functions of Random Variables, and Joint Distributions

8.1 The Poisson Distribution: The Last Discrete Distribution

The Poisson distribution counts the *number of events occurring in a fixed interval* (of time, space, area, ...).

Definition 8.1 (Poisson Distribution). $X \sim \text{Poisson}(\lambda)$, where $\lambda > 0$ is the **rate**: the average number of events per interval. Its PMF is

$$\mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x \in R_X = \{0, 1, 2, 3, \dots\}.$$

The value of X is a non-negative integer, so in particular it *can be zero*.

Proposition 8.1 (The Poisson PMF Is Well-Defined). $\sum_{x \in R_X} \mathbb{P}(X = x) = 1$.

Proof. Non-negativity is clear. For the total, use the Taylor series $e^\lambda = \sum_{x=0}^{\infty} \lambda^x / x!$:

$$\sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

■

Remark 8.1 (Understanding the Poisson: Rate, Shape, and Key Properties). **What λ really encodes.** The single parameter $\lambda > 0$ is simultaneously the *rate* of the process, the *expected*

count in one interval, and — remarkably — the *variance* of that count:

$$\mathbb{E}(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$

This equality of mean and variance is the Poisson's fingerprint: if you observe count data from the real world and find $\widehat{\mathbb{E}} \approx \widehat{\mathrm{Var}}$, the Poisson is the first model to try. (Data with $\mathrm{Var} \gg \mathbb{E}$ is called *overdispersed* and requires richer models such as the Negative Binomial.)

The three Poisson conditions. Underlying the definition is a physical story. A process generates Poisson(λ) counts in a fixed interval whenever:

(i) **Stationarity.** The probability of an event in any sub-interval depends only on its *length*, not on where in time it sits.

(ii) **Independence.** Events in non-overlapping sub-intervals are independent of one another.

(iii) **Rarity (no simultaneous events).** The probability of two or more events in a sub-interval of length h is $o(h)$ — negligible compared to h as $h \to 0$.

The Poisson Approximation Theorem formalises this: chunk the interval into n tiny pieces each of length $1/n$; each piece independently has one event with probability $p = \lambda/n$; the total count is Bin($n, \lambda/n$); send $n \to \infty$ and the Binomial converges to Poisson(λ).

Shape of the distribution. For small λ (say $\lambda = 1$), the distribution is strongly right-skewed: most of the mass sits at 0 and 1, with a long thin tail. As λ grows the distribution becomes more symmetric and bell-shaped, and by the Central Limit Theorem (which we will prove later) Poisson(λ) converges to $\mathcal{N}(\lambda, \lambda)$ as $\lambda \to \infty$ — another instance of the universal pull toward the Gaussian.

Superposition and splitting. Two of the most useful properties of the Poisson are closure under addition and thinning:

- **Superposition.** If $X \sim$ Poisson(λ_1) and $Y \sim$ Poisson(λ_2) are independent, then $X + Y \sim$ Poisson($\lambda_1 + \lambda_2$). Merge two independent Poisson streams and you get another Poisson stream, at the combined rate. (Think: ER arrivals by ambulance at rate 2/hr plus walk-in arrivals at rate 5/hr give total arrivals at rate 7/hr, still Poisson.)
- **Thinning.** If each event is independently kept with probability q, the thinned count is Poisson($q\lambda$). A quality-control inspector who flags each defect with probability q sees a Poisson($q\lambda$) process of flagged defects.

Real-world Poisson phenomena. The conditions above are satisfied (approximately) in a remarkable variety of settings:

- Number of radioactive decays from a sample in one second.
- Number of meteorites striking Earth per year of a given size range.
- Number of typos per page in a typeset book.
- Number of calls arriving at a call centre per minute.
- Number of mutations in a strand of DNA per replication.
- Number of goals scored in a soccer match (the reason the Poisson is beloved in sports analytics).

In each case, events are rare relative to the opportunity set, occur independently, and the rate is roughly stable over the interval — exactly the three conditions above.

When do we use the Poisson?

(1) As an *approximation to the binomial* when $n \to \infty$ and $p \to 0$ (Theorem below): whenever you have many independent trials each with a tiny success probability, the exact Binomial is

unwieldy but the Poisson with $\lambda = np$ is simple and accurate.

(2) To *model occurrences of rare events* in continuous time or space: whenever events arrive at a roughly constant rate, independently of one another, with negligible chance of two events at the same instant.

Theorem 8.2 (Poisson Approximation of the Binomial). *Let $X \sim \text{Bin}(n, p)$ with $n \to \infty$, $p \to 0$, and $np \to \lambda$ (a constant). Then for each fixed x,*

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \xrightarrow[n\to\infty,\ np\to\lambda]{} e^{-\lambda}\frac{\lambda^x}{x!}.$$

In practice: n large, p small $\implies$ $\text{Bin}(n, p) \approx \text{Poisson}(np)$.

Proof sketch. Substitute $p = \lambda/n$:

$$\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \underbrace{\frac{n(n-1)\cdots(n-x+1)}{n^x}}_{\to 1}\cdot\frac{\lambda^x}{x!}\cdot\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}}\cdot\underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\to 1} \longrightarrow \frac{\lambda^x}{x!}e^{-\lambda}. \quad \blacksquare$$

Remark 8.2 (Chunking Time: Why Counts in an Interval Are Poisson). We are at Boston University, so let us use a place nearby. Consider the chance of a car crash at Kenmore Square over one month. *Chunk* the month into n tiny sub-intervals (say, seconds). For large n, each sub-interval contains at most one crash, crashes in different sub-intervals are (approximately) independent, and each sub-interval carries the same tiny probability $p = \lambda/n$, where λ is the average number of crashes per month. The total count is then $\text{Bin}(n, \lambda/n)$, and letting the chunks shrink ($n \to \infty$) the theorem above turns this into exactly $\text{Poisson}(\lambda)$. Typical Poisson models: the number of accidents at Kenmore Square in a month; the number of ER arrivals between 5 and 6 pm.

Example 8.1 (ER Arrivals with Rate $\lambda = 6.9$). Suppose $\hat{\lambda} = 6.9$ arrivals per hour, so assume $X \sim \text{Poisson}(6.9)$ for the 5–6 pm hour. Find the probability that, in this hour,

(a) exactly 4 patients arrive;
(b) at least 2 patients arrive;
(c) between 4 and 10 patients (inclusive) arrive.

Solution. Throughout, $e^{-6.9} \approx 0.001008$.

(a) Direct plug-in:

$$\mathbb{P}(X = 4) = e^{-6.9}\frac{6.9^4}{4!} = e^{-6.9}\cdot\frac{2266.7}{24} \approx 0.095.$$

(b) The support is infinite, so "at least 2" calls for the complement rule:

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - e^{-6.9}(1 + 6.9) \approx 1 - 0.008 = 0.992.$$

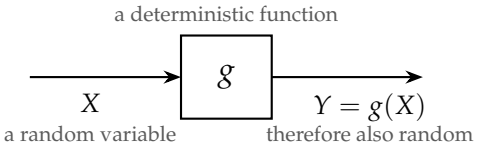
(c) A finite window: sum the PMF term by term,

$$\mathbb{P}(4 \leq X \leq 10) = \sum_{x=4}^{10} e^{-6.9}\frac{6.9^x}{x!} = \mathbb{P}(4) + \mathbb{P}(5) + \cdots + \mathbb{P}(10) \approx 0.821.$$

$\blacksquare$

8.2 Functions of Random Variables

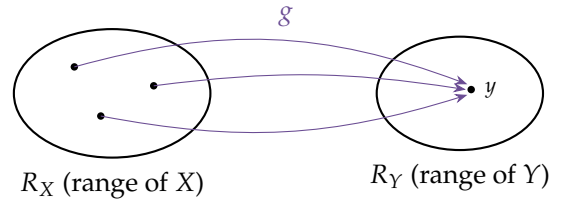
A random variable can be *passed through* a deterministic function: $Y = g(X)$. Since X is random and g is applied to whatever value X takes, the output Y is *also* a random variable.



Theorem 8.3 (PMF of a Function of a Discrete Random Variable)**.** *Let X be discrete and $Y = g(X)$. The range of Y is $R_Y = \{g(x) : x \in R_X\}$, and for each $y \in R_Y$,*

$$\mathbb{P}(Y = y) = \sum_{x : g(x) = y} \mathbb{P}(X = x)$$

— collect all *x-values that g maps to y, and add their probabilities.*



several x's can land on the same y — sum their probabilities

Example 8.2 (A Game with Payoff $Y = X^2$)**.** A student plays a game whose outcome is

$$X = \begin{cases} +1 & \text{win, with probability } 1/3, \\ 0 & \text{draw, with probability } 1/3, \\ -1 & \text{lose, with probability } 1/3. \end{cases}$$

Let $Y = g(X) = X^2$. Find the PMF of Y.

Solution. Squaring the three possible values gives $(\pm 1)^2 = 1$ and $0^2 = 0$, so $R_Y = \{0, 1\}$. Collect the x's behind each y:

$$\mathbb{P}(Y = 0) = \mathbb{P}\big(X^2 = 0\big) = \mathbb{P}(X = 0) = \frac{1}{3},$$
$$\mathbb{P}(Y = 1) = \mathbb{P}\big(X^2 = 1\big) = \mathbb{P}(X = +1) + \mathbb{P}(X = -1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

The two values $x = \pm 1$ merge into the single value $y = 1$, so their probabilities add. (Check: $\frac{1}{3} + \frac{2}{3} = 1$.) ■

Example 8.3 (Poisson Through an Absolute Value: $U = |X - 3|$)**.** Let $X \sim \text{Poisson}(\lambda)$ and define $U = |X - 3|$. Find the PMF of U.

Solution. The range of X is $\{0,1,2,\dots\}$; applying $x \mapsto |x-3|$ gives $R_U = \{0,1,2,3,\dots\}$. For each u, the solutions of $|x-3| = u$ are $x = 3-u$ and $x = 3+u$ (the first exists only when $3-u \geq 0$):

$$\begin{aligned}
\mathbb{P}(U=0) &= \mathbb{P}(X=3) = e^{-\lambda}\frac{\lambda^3}{3!},\\
\mathbb{P}(U=1) &= \mathbb{P}(X=2)+\mathbb{P}(X=4) = e^{-\lambda}\Big(\frac{\lambda^2}{2!}+\frac{\lambda^4}{4!}\Big),\\
\mathbb{P}(U=2) &= \mathbb{P}(X=1)+\mathbb{P}(X=5) = e^{-\lambda}\Big(\frac{\lambda^1}{1!}+\frac{\lambda^5}{5!}\Big),\\
\mathbb{P}(U=3) &= \mathbb{P}(X=0)+\mathbb{P}(X=6) = e^{-\lambda}\Big(1+\frac{\lambda^6}{6!}\Big),
\end{aligned}$$

and for $u \geq 4$ only the right branch $x = 3+u$ survives:

$$\mathbb{P}(U=u) = \mathbb{P}(X=3+u) = e^{-\lambda}\frac{\lambda^{3+u}}{(3+u)!}, \qquad u = 4,5,6,\dots$$

■

8.3 Jointly Discrete Random Variables

One experiment can carry *two or more* random variables at once:

- the temperature *and* pressure of a system;
- the amount of rain in two cities;
- for a language model, the input *and* the output.

Definition 8.2 (Joint PMF). Let X and Y be discrete random variables on the same experiment, with ranges R_X and R_Y. The **joint probability mass function** of the pair (X,Y) is

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

A valid joint PMF must satisfy

(i) $p_{X,Y}(x,y) \geq 0$ for all (x,y);
(ii) $\displaystyle\sum_{x\in R_X}\sum_{y\in R_Y} p_{X,Y}(x,y) = 1$.

8.4 Marginal Distributions

From the joint PMF we recover the distribution of each variable alone: fix one value and sum over *all* values of the other variable.

Definition 8.3 (Marginal PMFs). Given the joint PMF $p_{X,Y}$,

$$\begin{aligned}
p_X(x) &= \mathbb{P}(X=x) = \sum_{y\in R_Y} p_{X,Y}(x,y) \quad \text{for } x \in R_X,\\
p_Y(y) &= \mathbb{P}(Y=y) = \sum_{x\in R_X} p_{X,Y}(x,y) \quad \text{for } y \in R_Y.
\end{aligned}$$

9 Lecture 9: July 13, 2026 — Joint Distributions Continued, Independent Random Variables, and Expectation

9.1 Joint Distributions, Continued

Example 9.1 (Urn with Three Balls: A Full Joint Table). An urn holds three balls, numbered $1, 2, 3$. Draw two balls *without replacement*. Let X be the number on the first ball and Y the number on the second, so $R_X = R_Y = \{1, 2, 3\}$. Find the joint PMF $p_{X,Y}(x, y)$ and both marginals.

Solution. The sample space of ordered pairs is

$$\Omega = \{(1,2),\ (1,3),\ (2,1),\ (2,3),\ (3,1),\ (3,2)\},$$

six outcomes, each equally likely, so $\mathbb{P}(X = x,\ Y = y) = \frac{1}{6}$ whenever $x \neq y$, and 0 when $x = y$ (impossible without replacement). Summing along rows and columns gives the marginals: each is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

$p_{X,Y}(x,y)$	$y=1$	$y=2$	$y=3$	$p_X(x)$
$x=1$	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
$x=2$	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{3}$
$x=3$	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{3}$
$p_Y(y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

The interior cells hold the *joint* PMF $\mathbb{P}(X = x,\ Y = y) = p_{X,Y}(x, y)$; the margins hold the *marginal* PMFs. Both X and Y are (marginally) uniform on $\{1, 2, 3\}$. ■

Definition 9.1 (Conditional PMF). For discrete random variables X, Y with $\mathbb{P}(Y = y) > 0$,

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x,\ Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

This is just the conditional probability of events, applied to $A = \{X = x\}$ and $B = \{Y = y\}$.

In the urn example, $\mathbb{P}(X = 1 \mid Y = 2) = \frac{1/6}{1/3} = \frac{1}{2}$: knowing that the second ball shows 2 leaves the first ball uniform on $\{1, 3\}$.

9.2 Independent Random Variables

Remark 9.1 (From Independent Events to Independent Random Variables). Recall the definition of independence of events: $A \perp B$ meant that knowing A occurred left our belief about B unchanged, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. A random variable X is nothing but a *function* $X : \Omega \to \mathbb{R}$ that reads off a number from each outcome $\omega \in \Omega$. So to say two random variables X and Y are independent, we are really asking for a whole *family* of events built from these functions to be independent of one another: for every set $A \subseteq \mathbb{R}$ the preimage event $\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$, and likewise $\{Y \in B\}$, and we want

$$\{X \in A\} \perp \{Y \in B\} \qquad \text{for } \textit{every} \text{ choice of } A, B.$$

One pair of events being independent is not enough — independence of the random variables themselves means *all* such pairs are simultaneously independent, no matter which sets A and B we probe with. (Rigorously, A and B should range over the *Borel sets* of $\mathbb{R}$ — the sets to which probability can consistently be assigned — rather than literally every subset of $\mathbb{R}$; this technicality is beyond the scope of this course, and for the discrete random variables we study it never matters in practice.)

Definition 9.2 (Independence of Random Variables)**.** Random variables X and Y are **independent**, written $X \perp Y$, if

$$\mathbb{P}(X \in A,\ Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B) \qquad \text{for all sets } A, B \subseteq \mathbb{R}.$$

For *discrete* random variables this is equivalent to the pointwise criterion

$$\mathbb{P}(X = x,\ Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) \qquad \text{for } \textit{all } x \in R_X,\ y \in R_Y,$$

i.e. *every* cell of the joint table equals the product of its margins.

Example 9.2 (Flipping Two Coins)**.** Flip two fair coins. Let X code the first coin and Y the second, each via $H \mapsto 0$, $T \mapsto 1$, so $R_X = R_Y = \{0, 1\}$. Are X and Y independent?

Solution. The four outcomes are equally likely, so the joint PMF is

$$\mathbb{P}(X = x,\ Y = y) = \tfrac{1}{4} \quad \text{for all } x, y \in \{0, 1\},$$

and the marginals are

$$\mathbb{P}(X = x) = \tfrac{1}{2}\ \ (x \in \{0, 1\}), \qquad \mathbb{P}(Y = y) = \tfrac{1}{2}\ \ (y \in \{0, 1\}).$$

Check the product in every cell:

$$\mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4} = \mathbb{P}(X = x,\ Y = y) \quad \text{for all } x, y \in \{0, 1\}.$$

Joint = product of marginals everywhere, hence the two coin flips are independent: $X \perp Y$. ■

Remark 9.2 (Equal Marginals Do Not Imply Independence)**.** The urn example from the beginning of this lecture is a perfect counterexample. There, both X and Y are (marginally) uniform on $\{1, 2, 3\}$, so every marginal probability equals $\frac{1}{3}$. If the product rule held, every cell of the joint table would contain $\frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$. But the joint table has 0s on the diagonal (you cannot draw the same ball twice without replacement): $\mathbb{P}(X = 1, Y = 1) = 0 \neq \frac{1}{9}$. One failing cell is enough to conclude $X \not\perp Y$.

The moral: to verify independence of random variables, you must check the product rule in *every* cell of the joint table, not just inspect the marginals. Conversely, to *disprove* independence, a single cell where $\mathbb{P}(X = x, Y = y) \neq \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$ suffices.

Why does sampling without replacement kill independence? When you draw the first ball and it shows x, the ball is gone — so the second draw is constrained to $\{1, 2, 3\} \setminus \{x\}$. Learning $X = x$ tells you Y cannot equal x, so the events $\{X = x\}$ and $\{Y = x\}$ are not

independent for any x. Sampling *with* replacement removes this constraint: each draw is a fresh copy of the urn, so what you saw first has no bearing on the second — that is exactly the Binomial setting, where trials are independent by construction.

9.3 Expectation

Definition 9.3 (Expectation of a Discrete Random Variable)**.**

$$\mathbb{E}(X) = \sum_{x \in R_X} x \cdot \mathbb{P}(X = x)\,,$$

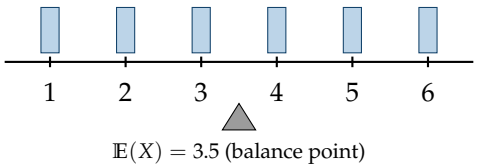
the **probability-weighted average** of all possible values.

Example 9.3 (Rolling a Fair Die)**.** Roll a fair die and let X be the number on the face, so $R_X = \{1, \ldots, 6\}$ and (discrete uniform) $\mathbb{P}(X = x) = \frac{1}{6}$ for each x. Find $\mathbb{E}(X)$.

Solution. All weights are equal, so the weighted average is the plain average:

$$\mathbb{E}(X) = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5.$$

"Just average them" works here only because the uniform PMF puts the same weight $\frac{1}{6}$ on every value; with unequal weights (e.g. the Bernoulli below) each value must be weighted by its own probability.



■

Remark 9.3 ($\mathbb{E}(X)$ Need Not Be a Possible Value)**.** $3.5 \notin R_X = \{1, \ldots, 6\}$, and that is perfectly fine. The expectation is the *centre of mass* of the distribution (where the PMF "balances"), not a value the random variable must actually take.

Example 9.4 (Bernoulli Expectation)**.** Let $X \sim \text{Bernoulli}(p)$: $R_X = \{0, 1\}$ with $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$. Then

$$\mathbb{E}(X) = \sum_{x \in \{0,1\}} x\, p_X(x) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Theorem 9.1 (Indicator Variables: Expectation = Probability)**.** *Let A be a random event, and*

*define the **indicator** of A:*

$$\mathbf{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Then

$$\mathbb{E}[\mathbf{1}_A] = 0 \cdot \mathbb{P}\left(A^C\right) + 1 \cdot \mathbb{P}(A) = \mathbb{P}(A)\,.$$

Theorem 9.2 (Linearity of Expectation). *For random variables $X_1, X_2, \ldots, X_n$ (on the same experiment),*

$$\mathbb{E}\big(X_1 + X_2 + \cdots + X_n\big) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n).$$

Example 9.5 (Binomial Expectation, Two Methods). Let $X \sim \text{Bin}(n, p)$. Find $\mathbb{E}(X)$.

Solution. **Method 1 (direct formula, combinatorics).** Grind through

$$\mathbb{E}(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

using the identity $x\binom{n}{x} = n\binom{n-1}{x-1}$; after re-indexing, the binomial theorem collapses the sum. It works, but it is heavy.

Method 2 (sum of Bernoullis). A binomial counts successes in n independent Bernoulli trials, so write

$$X = \sum_{i=1}^{n} Z_i, \qquad Z_i \sim \text{Bernoulli}(p) \text{ independent}, \quad Z_i = \mathbf{1}_{\{\text{trial } i \text{ succeeds}\}}.$$

By linearity of expectation and $\mathbb{E}(Z_i) = p$,

$$\mathbb{E}(X) = \mathbb{E}\Big(\sum_{i=1}^{n} Z_i\Big) = \sum_{i=1}^{n} \mathbb{E}(Z_i) = \sum_{i=1}^{n} p = n\,p.$$

The indicator decomposition is the workhorse here, and we shall reuse it repeatedly. ■

Example 9.6 (Poisson Expectation). Let $X \sim \text{Poisson}(\lambda)$: PMF $\mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $R_X = \{0, 1, 2, \ldots\}$. Find $\mathbb{E}(X)$.

Solution. The $x = 0$ term contributes nothing; then cancel one factor of x into $x!$ and pull out one λ:

$$\mathbb{E}(X) = \sum_{x=0}^{\infty} x\, e^{-\lambda}\frac{\lambda^x}{x!} = \sum_{x=1}^{\infty} e^{-\lambda}\frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \overset{k=x-1}{=} \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\, e^{\lambda} = \lambda.$$

The last step is the Taylor series $\sum_{k\geq 0} \lambda^k/k! = e^{\lambda}$. ■

9.4 Table of Expectations

Distribution	Parameters	Expectation $\mathbb{E}(X)$
Bernoulli	p	p
Binomial	n, p	$n\,p$
Hypergeometric	N, M, n	$n \cdot \dfrac{M}{N}$ (i.e. np with $p = \frac{M}{N}$)
Poisson	λ	λ
Geometric	p	$\dfrac{1}{p}$
Negative binomial	r, p	$\dfrac{r}{p}$

10 Lecture 10: July 14, 2026 — Expectation of a Function, Properties of Expectation, and Variance

10.1 Expectation of a Function of a Random Variable

A random variable X models an uncertain quantity. Once X is defined, every deterministic function $g : \mathbb{R} \to \mathbb{R}$ applied to X produces a *new* random variable $g(X)$: if X is the outcome of a die roll, then X^2 is its square, $\ln X$ is its logarithm, $\mathbf{1}_{X>3}$ is the indicator that it exceeded 3. Each of these is itself a random variable, and we may legitimately ask for its expectation.

The naive approach would be: first derive the PMF of $Y = g(X)$ from scratch (collect which values of x map to each value $y = g(x)$, sum their probabilities), and then apply the definition $\mathbb{E}(Y) = \sum_y y\, p_Y(y)$. This works, but it is laborious — you have to rebuild a new probability table for every new function g.

The **Law of the Unconscious Statistician (LOTUS)** is the observation that you do not need the PMF of $g(X)$ at all. You can compute $\mathbb{E}[g(X)]$ directly by summing $g(x)$ against the *original* PMF of X:

$$\mathbb{E}\big[g(X)\big] = \sum_{x \in R_X} g(x)\, p_X(x).$$

The name is self-deprecating: it is the formula that a statistician might write down "unconsciously", almost without thinking, and it turns out to be correct. The mathematical content is that $\mathbb{E}$ is an *integral* (here a weighted sum), and changing variables inside an integral by the function g is exactly what the change-of-variables formula does. In the discrete setting the proof is a clean regrouping of sums, which we carry out below.

Why does this matter? Variance — the central object of the next section — is defined as $\mathbb{E}[(X-\mu)^2]$. That is precisely $\mathbb{E}[g(X)]$ with $g(x) = (x-\mu)^2$, a non-linear function of X. Without LOTUS you would need to re-derive the PMF of $(X-\mu)^2$ every time. With LOTUS you just plug in. The same shortcut powers every moment calculation in this course: $\mathbb{E}[X^2]$, $\mathbb{E}[X^3]$, $\mathbb{E}[e^{tX}]$ (the moment-generating function, met in future courses) — all are instances of the same one-line formula.

Theorem 10.1 (Expectation of $g(X)$ — LOTUS)**.** *For a discrete random variable X and a function g,*

$$\mathbb{E}\big[g(X)\big] = \sum_{x \in R_X} g(x)\, p_X(x).$$

More generally, for two jointly distributed variables,

$$\mathbb{E}\big[g(X_1, X_2)\big] = \sum_{x_1 \in R_{X_1}} \sum_{x_2 \in R_{X_2}} g(x_1, x_2)\, \mathbb{P}(X_1 = x_1,\, X_2 = x_2)\,.$$

Remark 10.1 (One Sum or Two: The Same Thing)**.** Writing $\sum_{x_1 \in R_{X_1},\, x_2 \in R_{X_2}}$ under a single Σ, or as the iterated double sum $\sum_{x_1 \in R_{X_1}} \sum_{x_2 \in R_{X_2}}$, makes no difference: both run over exactly the same pairs (x_1, x_2), and for the finite (or absolutely convergent) sums we work with, the order of summation is immaterial.

Proof. Let $Y = g(X)$. By the definition of expectation, and then the PMF of a function of a random variable (Lecture 8: $\mathbb{P}(Y = y) = \sum_{x:\, g(x) = y} \mathbb{P}(X = x)$):

$$\mathbb{E}(Y) = \sum_{y \in R_Y} y\, \mathbb{P}(Y = y) = \sum_{y \in R_Y} y \sum_{\substack{x \in R_X \\ g(x) = y}} \mathbb{P}(X = x) = \sum_{y \in R_Y} \sum_{\substack{x \in R_X \\ g(x) = y}} g(x)\, \mathbb{P}(X = x)\,,$$

where in the last step $y = g(x)$ for every x in the inner sum. The sets $\{x : g(x) = y\}$ over the different $y \in R_Y$ split R_X into disjoint groups, and every $x \in R_X$ belongs to exactly one group, so the double sum is merely a regrouped version of the single sum:

$$\mathbb{E}(Y) = \sum_{x \in R_X} g(x)\, \mathbb{P}(X = x)\,. \qquad \blacksquare$$

10.2 Properties of Expectation

Theorem 10.2 (Properties of Expectation)**.** *For random variables X, Y and a constant c:*

(i) ***Scaling:*** $\mathbb{E}[cX] = c\,\mathbb{E}[X]$.
(ii) ***Constants:*** $\mathbb{E}[c] = c$.
(iii) ***Additivity:*** $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Example 10.1 (Linearity in Action)**.** Given random variables X, Y, Z, simplify $\mathbb{E}\big[\,3X + 4Y - 6Z - 2\,\big]$.

Solution. Combining all three properties,

$$\mathbb{E}\big[3X + 4Y - 6Z - 2\big] = 3\,\mathbb{E}[X] + 4\,\mathbb{E}[Y] - 6\,\mathbb{E}[Z] - 2.$$

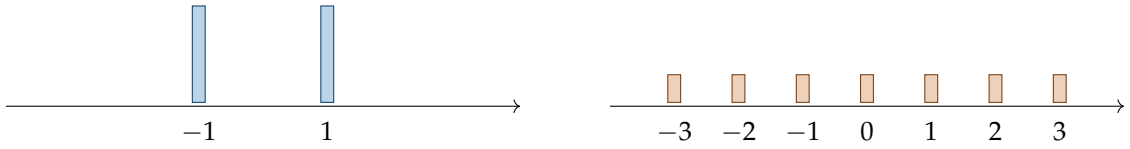
$\blacksquare$

10.3 Variance: Measuring Spread

Example 10.2 (Two Distributions with the Same Mean). Compare

$$X: \quad R_X = \{-1, 1\}, \quad \mathbb{P}(X = x) = \tfrac{1}{2}, \qquad\qquad Y: \quad R_Y = \{-3, -2, -1, 0, 1, 2, 3\}, \quad \mathbb{P}(Y = y) = \tfrac{1}{7}.$$

Both are symmetric around 0, so $\mathbb{E}(X) = 0 = \mathbb{E}(Y)$:



X: two values, weight $\frac{1}{2}$ each $\qquad$ Y: seven values, weight $\frac{1}{7}$ each

Yet Y is clearly more spread out than X. **Expectation is only one descriptor, and it is not enough.** We need a second number that measures spread around the mean.

Remark 10.2 (Why One Number Is Never Enough). Think of the expectation as the answer to the question: *"where does this random variable tend to land?"* It is a single real number that collapses an entire distribution down to its centre of mass. For many purposes that one-number summary is indispensable — but it is also *brutally lossy*. The example above makes this precise: X and Y have identical expectations, yet any reasonable notion of "behaviour" separates them instantly. X never strays further than 1 from 0; Y regularly lands at ± 3. If you only knew $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, you could not distinguish a calm random variable from a wild one.

More formally: the expectation captures *one moment* of the distribution — the first. But a distribution is, in general, an infinite object. Even two moments (mean and variance) leave infinitely many distributions indistinguishable; three moments do better; in the limit you need the entire distribution. The practical lesson is that *summarising randomness is always a choice about what you are willing to forget*, and you should always ask whether what you forgot matters for your problem.

Here is what we do keep with variance. Rather than asking "where does X land on average?", we ask: *"how far from its own mean does X tend to stray?"* A natural candidate would be the average deviation $\mathbb{E}[X - \mu]$, but that is always exactly zero (positives and negatives cancel, by linearity). So we square first to make all deviations positive, then average:

$$\mathrm{Var}(X) = \mathbb{E}\left[\,(X - \mu)^2\,\right].$$

Squaring penalises large deviations *disproportionately* — a point that is 2σ away contributes four times as much as a point that is σ away. This is both the strength of variance (it is sensitive to outliers and spread) and its limitation (a single extreme outcome can dominate the whole quantity). The choice to square is a modelling decision, not a logical necessity — but it leads to beautiful algebra, and that is largely why it won.

Definition 10.1 (Variance).

$$\mathrm{Var}(X) \overset{\text{def}}{=} \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)^2\Big] \qquad \textbf{Shortcut:} \quad \mathrm{Var}(X) = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2.$$

Variance is the expected *squared* distance from the mean: always ≥ 0, and large when the distribution puts weight far from $\mathbb{E}(X)$.

Proof of the shortcut. Write $\mu = \mathbb{E}(X)$ (a constant), expand the square, and use linearity:

$$\mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] = \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \quad \blacksquare$$

11 Lecture 11: July 16, 2026 — Variance Continued, Standard Deviation, and Covariance

(Wednesday, July 15 was Test 1 — no lecture.)

11.1 Variance, Continued

Example 11.1 (Bernoulli Variance). Let $X \sim \text{Bernoulli}(p)$, so $\mathbb{E}[X] = p$. Find $\text{Var}(X)$.

Solution. Since X takes only the values 0 and 1, we have $X^2 = X$, so $\mathbb{E}[X^2] = \mathbb{E}[X] = p$. By the shortcut formula,

$$\text{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = p - p^2 = p\,(1-p) = p\,q,$$

writing $q = 1 - p$ for the failure probability. $\blacksquare$

Theorem 11.1 (Properties of Variance). *For a random variable X and a constant c:*

$$\text{Var}(c) = 0, \qquad \text{Var}(cX) = c^2\,\text{Var}(X), \qquad \text{Var}(c+X) = \text{Var}(X).$$

Remark 11.1 (Shifting Moves the Mean, Not the Spread). $\text{Var}(c+X) = \text{Var}(X)$: if the Harvard stipend is c dollars more than the Boston University one, the *spread* of stipends is unchanged; the whole distribution just slides over by c. Scaling is different: $\text{Var}(cX) = c^2\,\text{Var}(X)$, because distances from the mean stretch by $|c|$ and variance squares them.

Proof. $\text{Var}(c) = 0$**:** a constant never deviates from its mean $\mathbb{E}(c) = c$,

$$\text{Var}(c) = \mathbb{E}\left[(c - \mathbb{E}(c))^2\right] = \mathbb{E}\left[(c-c)^2\right] = \mathbb{E}[0] = 0.$$

$\text{Var}(cX) = c^2\,\text{Var}(X)$**:** with $\mu = \mathbb{E}(X)$, $\mathbb{E}(cX) = c\mu$, so

$$\text{Var}(cX) = \mathbb{E}\left[(cX - c\mu)^2\right] = \mathbb{E}\left[c^2(X-\mu)^2\right] = c^2\,\mathbb{E}\left[(X-\mu)^2\right] = c^2\,\text{Var}(X).$$

$\text{Var}(c+X) = \text{Var}(X)$**:** $\mathbb{E}(c+X) = c+\mu$, and the c's cancel inside the square, $\big((c+X) - (c+\mu)\big)^2 = (X-\mu)^2$. $\blacksquare$

11.2 Standard Deviation

Definition 11.1 (Standard Deviation).

$$\sigma_X \stackrel{\text{def}}{=} \sqrt{\operatorname{Var}(X)} = \sqrt{\mathbb{E}\left[(X - \mathbb{E}(X))^2\right]}.$$

It carries the same information as the variance but lives on the *same scale* as X itself (variance carries squared units).

Example 11.2 (Bernoulli Standard Deviation). For $X \sim \text{Bernoulli}(p)$, $\sigma_X = \sqrt{\operatorname{Var}(X)} = \sqrt{p\,(1-p)}$. For a fair coin, $p = \frac{1}{2}$, so

$$\sigma_X = \sqrt{\tfrac{1}{2} \cdot \left(1 - \tfrac{1}{2}\right)} = \sqrt{\tfrac{1}{4}} = \tfrac{1}{2}.$$

Example 11.3 (Standard Deviation of a Linear Transform). Express σ_{3X-6} in terms of σ_X.

Solution. Shift does nothing, scale comes out squared, then the square root brings it back down:

$$\sigma_{3X-6} = \sqrt{\operatorname{Var}(3X-6)} = \sqrt{9\,\operatorname{Var}(X)} = 3\,\sigma_X.$$

■

Remark 11.2 (Variance, Risk, and Rare Events — From Wall Street to Three Sigma). **Variance as risk.** In finance, a random variable X often represents the return on an asset over some time horizon. The expected value $\mathbb{E}(X)$ is the anticipated gain; the standard deviation σ_X is the canonical measure of *risk* — how wildly the actual return might deviate from that expectation. A bond with $\sigma_X \approx 0$ is nearly certain to pay its coupon; a small-cap stock with large σ_X might double or halve. Portfolio managers routinely trade off $\mathbb{E}(X)$ against σ_X, a framework formalised by Harry Markowitz's *Modern Portfolio Theory* (Nobel Prize, 1990): given a target expected return, find the portfolio that minimises variance.

Two-sigma and three-sigma events. Once we reach the Central Limit Theorem later in the course, we will see that many natural averages follow (approximately) a Normal distribution $\mathcal{N}(\mu, \sigma^2)$. For a normal random variable, roughly:

- $\approx 95.4\%$ of outcomes fall within $\pm 2\sigma$ of the mean — a *two-sigma event* (a move beyond 2σ) occurs only about 4.6% of the time, or roughly once every three weeks of daily trading.
- $\approx 99.7\%$ of outcomes fall within $\pm 3\sigma$ — a *three-sigma event* occurs only about 0.3% of the time, or roughly once every 370 trading days.

These probabilities give risk managers a language for rare but plausible disasters. The infamous "Value at Risk" (VaR) models that banks use before the 2008 financial crisis treated losses beyond 3σ as essentially impossible — an assumption the crisis proved catastrophically wrong. (Financial returns have heavier tails than the normal distribution, so genuine 5σ or 6σ moves occur far more often than the Gaussian model predicts; this mismatch is called *tail risk*.)

Two Sigma — a hedge fund named after this idea. In 2001, David Siegel and John Overdeck founded *Two Sigma Investments* in New York, naming the firm directly after the two-standard-deviation threshold: their mandate was to find return signals that are statistically significant

— at least two sigma away from chance. Using systematic, data-driven strategies built on probability, statistics, and machine learning, Two Sigma has grown into one of the largest quantitative hedge funds in the world, managing over $60 billion in assets as of the mid-2020s. Their flagship Compass fund has delivered annualised returns in the range of 10–15% over a decade, with carefully controlled volatility — precisely by engineering portfolios whose σ is understood and budgeted rather than left to chance. The name is a daily reminder that rigorous statistical thinking, not intuition, is the product they sell.

The three-sigma principle in manufacturing. Beyond finance, the *three-sigma rule* (also called *Six Sigma* when applied symmetrically to both sides) is the backbone of quality control in manufacturing. If a factory targets a bolt diameter of μ mm with process standard deviation σ, then keeping the process within $\pm 3\sigma$ means at most 0.3% of bolts are defective — acceptable for most industries. "Six Sigma" processes push the threshold to $\pm 6\sigma$, reducing the defect rate to roughly 3.4 per million. The same arithmetic you learned today — σ, deviations from the mean, tail probabilities — powers global supply chains, pharmaceutical manufacturing, and semiconductor fabrication.

11.3 Standardisation

Definition 11.2 (Standardisation of a Random Variable)**.** The **standardisation** of X is

$$X^* = \frac{X - \mathbb{E}(X)}{\sigma_X} = \frac{X - \mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}},$$

i.e. subtract the mean, then divide by the standard deviation. The result always satisfies

$$\mathbb{E}\left(X^*\right) = 0, \qquad \mathrm{Var}\left(X^*\right) = 1.$$

Proof. Write $\mu = \mathbb{E}(X), \sigma = \sigma_X$. By linearity, $\mathbb{E}(X^*) = \frac{1}{\sigma}\big(\mathbb{E}(X) - \mu\big) = 0$; by the variance properties (shift does nothing, scale $\frac{1}{\sigma}$ comes out squared), $\mathrm{Var}(X^*) = \frac{1}{\sigma^2}\,\mathrm{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1$. ■

Remark 11.3 (Standardisation in the Wild — and a Glimpse of Things to Come)**.** Standardisation is one of those ideas that looks almost trivially simple — subtract a number, divide by another — yet it appears *everywhere* in modern data science and deep learning. A few highlights to broaden the horizon:

1. **Feature scaling in machine learning.** Before training a model, every input feature is typically standardised: height (cm), income ($), and age (years) all land on a common scale with mean 0 and standard deviation 1. Without this, gradient descent chases the feature with the largest raw variance and ignores the rest.
2. **Batch normalisation in neural networks.** Inside a deep neural network, after each hidden layer the activations are standardised *across the current batch of training examples*. This stabilises the distribution that each layer sees, allowing much deeper networks to be trained reliably — one of the key tricks behind modern large language models.
3. **Layer normalisation in Transformers.** The architecture behind GPT, BERT, and their descendants standardises activations *across the features of a single example* rather than across the batch. Same formula (X^*), different axis. This is what lets Transformers train stably on variable-length sequences.

4. **Standardising test scores (z-scores).** When comparing students across exams with different average difficulties, the raw score is meaningless; the standardised score (z-score) says "how many standard deviations above the class average?" — a fair, scale-free comparison.

A teaser for things to come. The most beautiful reason to care about standardisation is a theorem you will meet later in this course: the **Central Limit Theorem (CLT)**. Suppose you observe n independent, identically distributed random variables $X_1, X_2, \ldots, X_n$ (think: n repeated measurements, each with the same mean μ and standard deviation σ). Form their sample mean

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Then the standardised sample mean

$$\bar{X}_n^* = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges, as $n \to \infty$, to a distribution that is the same *regardless of what distribution the X_i originally followed*: the **standard normal (Gaussian)**, the famous bell curve $\mathcal{N}(0,1)$. The CLT is the reason the normal distribution appears throughout statistics, science, and engineering — and standardisation is the exact operation that reveals it.

11.4 Covariance: A First Look at Two Dimensions

So far everything has been one-dimensional. Moving to *two* random variables, the first question is how they vary *together*.

Every concept we have built — expectation, variance, standard deviation — describes a single random variable in isolation. But the world is rarely one-dimensional. A doctor records both blood pressure and cholesterol. An engineer monitors temperature and pressure simultaneously. A portfolio manager holds two assets whose prices move in the same market. In each case the *relationship* between the variables is at least as important as their individual behaviour.

Variance answered: *how far does X stray from its own mean, on average?* Covariance asks the next, two-dimensional question: *when X is above its mean, does Y tend to be above its mean too — or below?* More precisely, consider the signed product of the two deviations from their respective means,

$$\big(X - \mathbb{E}(X)\big) \cdot \big(Y - \mathbb{E}(Y)\big).$$

This product is *positive* when both deviations share the same sign (both above or both below their means), and *negative* when the deviations point in opposite directions. Averaging this product over all outcomes yields the covariance:

- $\mathrm{Cov}(X,Y) > 0$: X and Y tend to move *together* — they are **positively correlated**.
- $\mathrm{Cov}(X,Y) < 0$: when one rises the other tends to fall — they are **negatively correlated**.
- $\mathrm{Cov}(X,Y) = 0$: the deviations are *uncorrelated* on average; in particular, independence implies zero covariance (the converse is *not* true in general — zero covariance does not guarantee independence).

A financial picture. Let X be tomorrow's return on a technology stock and Y be tomorrow's return on an oil company. Intuitively: tech and oil respond to different economic forces, so their returns might move fairly independently ($\mathrm{Cov} \approx 0$). Now compare this to holding two tech stocks: both are sensitive to the same sector news, so their returns tend to rise and fall together ($\mathrm{Cov} > 0$). The sign of $\mathrm{Cov}(X,Y)$ therefore determines whether combining two assets amplifies or dampens

total risk — a consequence we will make precise once we have established the key properties of covariance below.

Definition 11.3 (Covariance).

$$\mathrm{Cov}(X,Y) = \mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)\cdot\big(Y-\mathbb{E}(Y)\big)\Big] \qquad \textbf{Shortcut: } \mathrm{Cov}(X,Y) = \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y].$$

Proof of the shortcut. Expand with $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$ and use linearity, exactly as for the variance shortcut:

$$\mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big] = \mathbb{E}[XY] - \mu_Y\mathbb{E}[X] - \mu_X\mathbb{E}[Y] + \mu_X\mu_Y = \mathbb{E}[XY] - \mu_X\mu_Y. \qquad \blacksquare$$

Theorem 11.2 (Properties of Covariance). *For random variables X, Y, Z and constants c, c_1, c_2:*

(i) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$;
(ii) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$;
(iii) $\mathrm{Cov}(cX,Y) = c\cdot\mathrm{Cov}(X,Y)$;
(iv) $\mathrm{Cov}(X,Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$;
(v) $\mathrm{Cov}(X,c) = 0, \quad \mathrm{Cov}(c_1,c_2) = 0.$

Remark 11.4 (Covariance, Independence, and the Algebra of Risk). **Independence kills covariance.** If $X \perp Y$ then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ (a consequence of independence we will prove shortly), so the shortcut formula gives $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$. Independence is a *sufficient* condition for zero covariance, but emphatically not *necessary*: one can construct random variables with $\mathrm{Cov}(X,Y) = 0$ that are heavily dependent. (A classic example: let $X \sim \mathrm{Uniform}\{-1,0,1\}$ and $Y = X^2$; then $\mathrm{Cov}(X,Y) = 0$ yet Y is a deterministic function of X.) Covariance detects only *linear* association; non-linear dependence can hide from it entirely.

Variance of a portfolio. Property (iv) (bilinearity) makes covariance the right tool for portfolio risk. Let $w_1, w_2 \geq 0$ with $w_1 + w_2 = 1$ be the weights invested in assets X and Y, and let $P = w_1X + w_2Y$ be the portfolio return. Then

$$\begin{aligned}\mathrm{Var}(P) &= \mathrm{Var}(w_1X + w_2Y)\\ &= w_1^2\ \mathrm{Var}(X) + 2\,w_1w_2\ \mathrm{Cov}(X,Y) + w_2^2\ \mathrm{Var}(Y).\end{aligned}$$

Three observations follow directly from this formula:

(1) If $\mathrm{Cov}(X,Y) = \sigma_X\sigma_Y > 0$ (perfect positive correlation), then $\mathrm{Var}(P) = (w_1\sigma_X + w_2\sigma_Y)^2$ — portfolio risk is the weighted average of individual risks with no reduction. Diversification achieves nothing.
(2) If $\mathrm{Cov}(X,Y) = 0$ (uncorrelated assets), then $\mathrm{Var}(P) = w_1^2\sigma_X^2 + w_2^2\sigma_Y^2 < w_1\sigma_X^2 + w_2\sigma_Y^2$: the portfolio variance is strictly less than the weighted average of the individual variances. Risk is reduced without sacrificing expected return.
(3) If $\mathrm{Cov}(X,Y) = -\sigma_X\sigma_Y$ (perfect negative correlation), then $\mathrm{Var}(P) = (w_1\sigma_X - w_2\sigma_Y)^2$, which equals *zero* when $w_1\sigma_X = w_2\sigma_Y$: a perfectly hedged portfolio with zero variance.

This is Markowitz's insight in equation form: *risk management is linear algebra applied to covariance*. A portfolio manager who holds tech, energy, and healthcare stocks is not just choosing three expected returns — they are choosing a position in a covariance matrix, hoping the off-diagonal terms are small (or negative) enough to reduce total variance. With n assets,

> the portfolio variance is $\mathbf{w}^\top \Sigma\, \mathbf{w}$, where Σ is the $n \times n$ covariance matrix — an object you will meet properly in linear algebra and multivariate statistics, but whose every entry is exactly $\mathrm{Cov}(X_i, X_j)$ as defined here.