Probability

An Undergraduate Textbook

Ying Nian Wu

Department of Statistics
University of California, Los Angeles

Based on the author's STATS 100A Lecture Slides
The book is written with the help of Claude 4
Some pictures are taken from the internet. Credits belong to original authors.

Preface

This book grew out of the lecture slides for STATS 100A, an undergraduate probability course taught in the Department of Statistics at UCLA. The goal of the course, and of this book, is to build a deep understanding of probability from the ground up, starting with the simple act of counting equally likely outcomes and arriving at the central limit theorem, stochastic processes, information theory, and connections to modern machine learning.

The organizing philosophy is captured in a single slogan: **as long as you can count**. Probability begins with counting a finite population of equally likely possibilities. From there, every major concept — events, random variables, conditional probability, independence, expectation, variance — emerges naturally. When we move to continuous random variables, we replace counting with measuring (lengths, areas, volumes), and the same logic applies. When we study repeated experiments, we count or measure within the *hyper-population of all possible sequences* of outcomes — the "N^n reasoning" that makes the law of large numbers an exercise in counting.

Throughout the book, we maintain two parallel tracks:

1. **Intuition, visualization, and motivation.** Every concept is introduced through concrete examples: rolling a die, sampling from a population, throwing points into a region, flipping coins, and random walks. Figures appear on nearly every page.

2. **Precise notation and formulas.** Every intuition is backed by a careful definition and a derivation. The two tracks reinforce each other: the intuition tells you *why* a formula is true, and the formula tells you *exactly what* the intuition means.

A distinctive feature of this book is its emphasis on connecting classical probability to modern applications. We discuss Bayes networks, GPT (Generative Pre-trained Transformers), diffusion models, deep learning, reinforcement learning, and information theory — not as separate "advanced topics," but as natural extensions of the basic counting and conditioning ideas developed in the first chapter.

The book is organized into four chapters:

- **Chapter 1: Basics and Examples.** A self-contained, example-driven tour of all the key concepts, from sample spaces to Bayes networks. This chapter is deliberately informal and builds intuition before formalism.

- **Chapter 2: Random Variables.** A systematic treatment of discrete and continuous random variables, including expectation, variance, and the major probability distributions (Bernoulli, Binomial, Geometric, Exponential, Normal).

- **Chapter 3: Two or More Random Variables.** Joint distributions, covariance, correlation, regression, independence, the law of large numbers, and the central limit theorem.

- **Chapter 4: Advanced Topics.** Continuous-time processes (Poisson, Brownian motion), proofs of normal approximation, transformations, simulation methods, the Jensen inequality, and information theory.

I hope this book helps you see that probability is not a collection of formulas to memorize, but a way of thinking about uncertainty that connects counting, geometry, physics, and computation.

Ying Nian Wu
Los Angeles, California

Contents

Chapter 1

Basics and Examples

This chapter is a self-contained tour of the fundamental ideas of probability, developed through six running examples. Our goal is to build strong intuition before we move to systematic treatments in later chapters. By the end of this chapter, you will have seen — informally but concretely — nearly every major concept in the course: sample spaces, events, probability, random variables, conditional probability, independence, the binomial distribution, the law of large numbers, Markov chains, and Bayes' rule.

The recurring theme throughout is a three-step pattern:

$$\textbf{Experiment} \ \rightarrow \ \textbf{Outcome} \ \rightarrow \ \textbf{Number}.$$

We perform an experiment (roll a die, sample a person, flip a coin), observe an outcome, and associate a number with that outcome. Probability is the study of how these numbers behave when the experiment involves randomness.

Think of this chapter as a guided hiking tour through probability's landscape. We will visit all the major landmarks quickly, getting a feel for the terrain, and then return in later chapters to explore each area in depth.

1.1 Sample Space and Events

1.1.1 Sample Space

Every probabilistic situation begins with an **experiment** whose result is uncertain. Before we perform the experiment, we do not know what will happen. After we perform it, we observe one specific result. The **sample space**, denoted Ω, is the set of all possible outcomes of the experiment. It is the complete catalog of everything that *could* happen.

Example 1.1.1 (Rolling a die)**.** Roll a single die. The sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

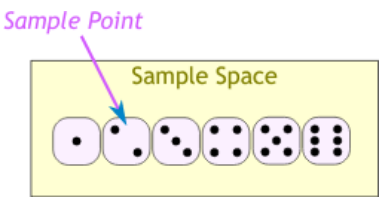


Figure 1.1: Rolling a die: six equally likely outcomes.

The key idea is: we **randomly sample an outcome from the sample space**. Before we roll the die, we do not know which face will come up. After we roll, we observe one specific element of Ω. The sample space tells us *what is possible*; probability will tell us *how likely* each possibility is.

Notice that the sample space depends on how we describe the experiment. If we only care about whether the number is even or odd, our sample space could be $\{\text{even}, \text{odd}\}$. But usually we prefer to work with the most detailed description available, because we can always combine outcomes later.

1.1.2 Events

An **event** is a statement about the outcome, or equivalently, a subset of the sample space. These two perspectives — statement and subset — are interchangeable, and switching freely between them is one of the most useful skills in probability.

Definition 1.1.1 (Event)**.** An event A is a subset $A \subseteq \Omega$. We say that event A *occurs* if the observed outcome ω belongs to A.

Why are events subsets? Because a statement like "the outcome is bigger than 4" picks out exactly those elements of Ω that make the statement true. The statement and the subset describe the same thing, just in different languages.

Example 1.1.2 (Die: event "bigger than 4")**.** Continuing Example 1.1.1, let A be the event that the number is bigger than 4.

1. As a **statement**: "the outcome is bigger than 4."
2. As a **subset**: $A = \{5, 6\}$.

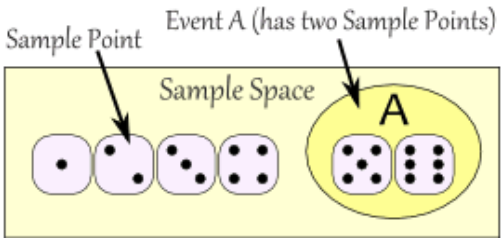


Figure 1.2: The event $A = \{5, 6\}$ highlighted among the six outcomes.

If we roll the die and get a 5, then the outcome $\omega = 5$ belongs to $A = \{5, 6\}$, so we say event A has occurred. If we get a 3, then $\omega = 3$ does not belong to A, so event A has not occurred.

1.1.3 Probability by Counting

Now comes the most natural question: how likely is event A to occur? If all outcomes in Ω are **equally likely** (the die is fair), then the answer is beautifully simple. The probability of A is the ratio of the size of A to the size of Ω:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{2}{6} = \frac{1}{3}.$$

Here $|A|$ counts the **size** of A, that is, the number of elements in A. This is perhaps the most natural definition of probability: the fraction of equally likely outcomes that belong to A.

Let us pause to appreciate how simple and powerful this idea is. To find a probability, we just need to do two things: count the total number of outcomes, and count how many of them belong to the event we care about. As long as every outcome is equally likely, the ratio gives us the probability. This is the essence of "as long as you can count."

1.1.4 Random Variables

So far, we have talked about outcomes (elements of Ω) and events (subsets of Ω). A **random variable** adds another layer: it is a number associated with the outcome. More precisely, it is a function from the sample space to the real numbers.

Definition 1.1.2 (Random variable)**.** A random variable X is a function $X : \Omega \to \mathbb{R}$. For each outcome $\omega \in \Omega$, $X(\omega)$ is a real number.

Why do we need random variables? Because in many situations, we do not care about the outcome itself but about some *numerical quantity* derived from it. If we roll two dice, we might care about the sum, the maximum, or the difference — each of these is a random variable.

Example 1.1.3 (Die: the number shown)**.** Let X be the number shown on the die. Then $X(\omega) = \omega$ for each $\omega \in \{1, 2, 3, 4, 5, 6\}$. The event "$X > 4$" is the same as the event $A = \{5, 6\}$, so

$$P(X > 4) = \frac{1}{3}.$$

In this example, the random variable is trivial — it just reports the outcome itself. But in more complex experiments, the random variable extracts a specific piece of information from a more complicated outcome. For instance, if the experiment is "sample a person from a population," then the outcome ω is a specific person, and random variables like $X(\omega) =$ height of person ω or $Y(\omega) =$ age of person ω extract numerical information from that person.

An event is a **mathematical statement about the random variable**. We can describe events using either the subset language ($A = \{5, 6\}$) or the random variable language ($X > 4$). We can use either events or random variables. In Parts 2 and 3, we will focus on random variables.

1.1.5 Conditional Probability

Suppose we learn that some event A has occurred. How does this change the probability of another event B? This is the idea of **conditional probability**: the probability of B *given* A. It is one of the most important concepts in all of probability.

Example 1.1.4 (Die: conditional probability)**.** Let $A = \{5, 6\}$ (bigger than 4) and let $B = \{6\}$ (equal to 6). Given that A happens — that is, given that the number is bigger than 4 — what is the probability of B?

Since we know the outcome is in $A = \{5, 6\}$, and $B = \{6\}$ is one of the two elements of A:

$$P(B \mid A) = \frac{1}{2}.$$

In random variable notation: $P(X = 6 \mid X > 4) = 1/2$.

Notice what happened: before we learned anything, the probability of rolling a 6 was $P(B) = 1/6$. But once we learned that the outcome is bigger than 4, the probability jumped to $P(B \mid A) = 1/2$. Learning new information changes probabilities.

The key intuition is that conditioning on A is **as if** we randomly sample from A. Once we know the outcome is in A, the event A becomes our new sample space. It is **as if** A is the sample space. Within this reduced sample space, the probability of B is the fraction of A that belongs to B:

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

This formula says: look at the outcomes that are in *both* A and B (that is, the intersection $A \cap B$), and divide by the total number of outcomes in A. We are doing exactly the same counting as before — "favorable divided by total" — but now the "total" is $|A|$ instead of $|\Omega|$ because we have restricted our attention to the sub-population A.

1.1.6 Set Operations on Events

Since events are subsets, we can combine them using set operations. Each set operation corresponds to a logical operation on statements. This connection between logic and set theory is one of the beautiful aspects of probability.

Complement. A^c is the event "not A." The statement is: not A. The subset is: $A^c = \{1, 2, 3, 4\}$. The complement consists of all outcomes that do *not* belong to A. If A is "the number is bigger than 4," then A^c is "the number is 4 or less."

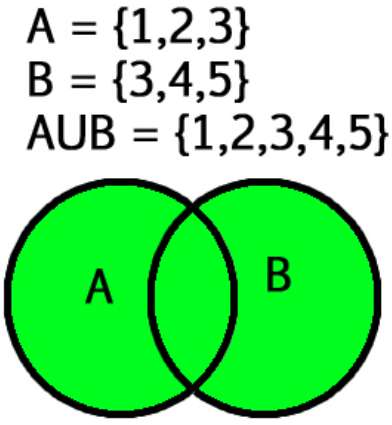


Figure 1.3: Venn diagram showing union $A \cup B$.

Union. $A \cup B$ is the event "A or B (or both)." Statement: A or B. Subset: $A \cup B$. The union collects all outcomes that belong to at least one of the two events. In everyday English, "or" sometimes means "one or the other but not both," but in mathematics, "or" always includes the possibility of both.

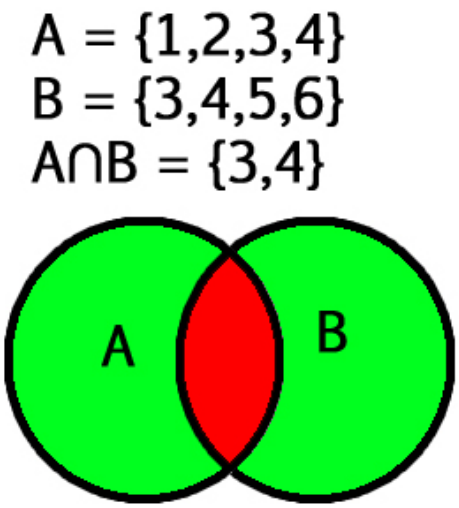


Figure 1.4: Venn diagram showing intersection $A \cap B$.

Intersection. $A \cap B$ is the event "A and B." Statement: A and B. Subset: $A \cap B$. The intersection consists of outcomes that belong to *both* events simultaneously. For example, if $A =$ "bigger than 4" and $B =$ "even," then $A \cap B = \{6\}$ because 6 is the only outcome that is both bigger than 4 and even.

1.2 The Sample Space as a Population

Our second major example shifts from a die to something more concrete and relatable: a population of people. This perspective is extremely powerful because it connects abstract probability to something tangible — counting people in a group.

Example 1.2.1 (Population sampling)**.** Consider a population of 100 people: 50 males and 50 females. Among the males, 30 are taller than 6 feet. Among the females, 10 are taller than 6 feet.

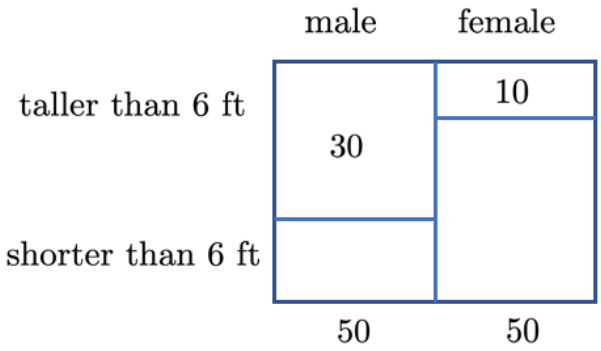


Figure 1.5: A population of 100 people, divided by gender and height.

The experiment is: **randomly sample one person from the population**. "Randomly" means each person is equally likely to be chosen — we might, for instance, put all 100 names in a

hat and draw one. The sample space Ω is the population itself — the set of all 100 people. Each person is an outcome, and each person is equally likely to be selected.

The sample space Ω is the population.

This identification — sample space equals population — is one of the most important conceptual moves in this book. Whenever you see a probability problem, try to think of a population of equally likely possibilities. This "population thinking" will serve you throughout the course.

1.2.1 Events as Sub-Populations

Let A be the event that the person is male, and B be the event that the person is taller than 6 feet. Then A is the **sub-population of males** (50 people), and B is the **sub-population of tall people** ($30 + 10 = 40$ people).

Every event corresponds to a sub-population, and every sub-population corresponds to an event. This gives us a very concrete way to think about events: they are simply groups of people within the larger population.

1.2.2 Probability as Population Proportion

Since all 100 people are equally likely to be selected:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{50}{100} = 50\%, \qquad P(B) = \frac{|B|}{|\Omega|} = \frac{30+10}{100} = 40\%.$$

This gives us an important interpretation: **probability equals population proportion**. The probability of being male is simply the proportion of the population that is male. The probability of being tall is the proportion of the population that is tall. There is nothing mysterious about probability here — it is just a fraction.

1.2.3 Conditional Probability as Sub-Population Proportion

Now consider conditional probabilities. The conditional probability $P(A \mid B)$ asks: among tall people, what is the proportion of males?

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{30}{40} = 75\%.$$

Here $A \cap B$ is the set of people who are both male and tall. There are 30 such people (the 30 tall males). Among the 40 tall people total, 30 are male, so the proportion is 75%.

Conversely, $P(B \mid A)$ asks: among males, what is the proportion of tall people?

$$P(B \mid A) = \frac{|A \cap B|}{|A|} = \frac{30}{50} = 60\%.$$

Among the 50 males, 30 are tall, so the proportion is 60%.

Notice that $P(A \mid B) = 75\%$ and $P(B \mid A) = 60\%$ — these are *different* numbers! This is an important point. Conditioning on B and asking about A is not the same as conditioning on A and asking about B. The proportion of males among tall people is not the same as the proportion of tall people among males. Confusing these two directions of conditional probability is one of the most common errors in probabilistic reasoning.

The key insight: **conditional probability equals proportion within a sub-population**. To compute $P(A \mid B)$, we restrict our attention to the sub-population B and ask what fraction of B belongs to A.

1.2.4 Random Variables as Functions of Outcomes

The link between events and random variables becomes very concrete in the population setting. Let $\omega \in \Omega$ be a person. Define:

- $X(\omega)$: the gender of person ω, coded as $X(\omega) = 1$ if ω is male, and $X(\omega) = 0$ if ω is female.
- $Y(\omega)$: the height of person ω (a real number).

Then the events become statements about random variables:

$$A = \{\omega : X(\omega) = 1\}, \qquad B = \{\omega : Y(\omega) > 6\}.$$

And the probabilities become:

$$\begin{aligned} P(A) &= P(\{\omega : X(\omega) = 1\}) = P(X = 1) = 50\%, \\ P(B) &= P(\{\omega : Y(\omega) > 6\}) = P(Y > 6) = 40\%, \\ P(B \mid A) &= P(Y > 6 \mid X = 1) = 60\%, \\ P(A \mid B) &= P(X = 1 \mid Y > 6) = 75\%. \end{aligned}$$

This is important: the random variables X and Y are simply functions that extract numerical information from each person in the population. The event $A = \{X = 1\}$ is the sub-population of males, and the event $B = \{Y > 6\}$ is the sub-population of tall people. Everything reduces to counting within a population.

1.2.5 Axiom 0

Whether we have a real population of people under purely random sampling, or an *imagined population of equally likely possibilities* (like the six faces of a die), the fundamental formula is the same:

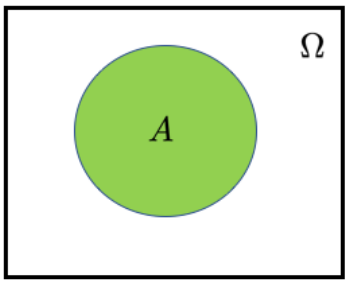


Figure 1.6: Probability as the ratio of favorable to total outcomes.

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

We call this **Axiom 0**. It is the starting point of all probability. We can always translate a problem into this equally likely setting. Whether the "population" is real (100 people) or imagined (six faces of a die), the logic is identical: probability is the fraction of the population that satisfies the condition.

1.2.6 Conditional Probability: Formal Definition

Starting from Axiom 0, we can derive the general formula for conditional probability. The trick is to divide both the numerator and the denominator by $|\Omega|$:

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{P(A \cap B)}{P(B)}.$$

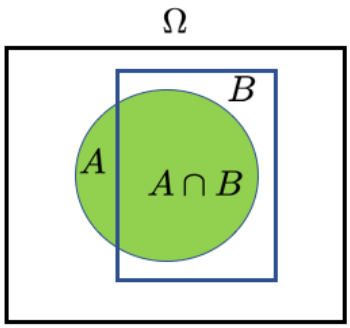


Figure 1.7: Conditional probability: $P(A \mid B) = P(A \cap B)/P(B)$.

This formula has two interpretations:

1. **Physical:** We sample from the sub-population B. The event B defines a *condition*, and we are asking for the probability of A within this restricted population.

2. **Mental:** We learn that B has happened, and update our beliefs accordingly. It is *as if* we sample from B.

We call this **Axiom 4** (or the definition of conditional probability). The physical interpretation says we literally restrict our world to B; the mental interpretation says we *imagine* restricting our world to B after learning that B occurred. Both lead to the same formula.

1.3 The Sample Space as a Region

Our third example moves from finite populations to continuous settings. This is a major conceptual leap, but the underlying idea is exactly the same. The key insight is that a region in space is a "population of points" — an uncountably infinite population. Just as we counted people in a finite population, we now *measure* regions of space.

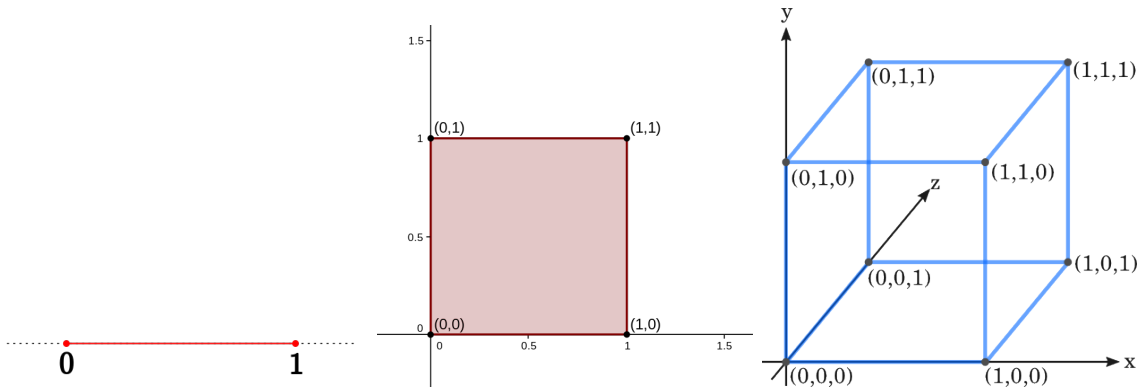


Figure 1.8: Left: X uniform on $[0,1]$. Center: (X,Y) uniform on $[0,1]^2$. Right: (X,Y,Z) uniform on $[0,1]^3$.

Consider three settings:

1. X is a uniform random number in $[0,1]$. The sample space is $\Omega = [0,1]$.

2. (X,Y) are two independent uniform random numbers in $[0,1]$. The sample space is $\Omega = [0,1]^2$, the unit square.

3. (X,Y,Z) are three independent uniform random numbers in $[0,1]$. The sample space is $\Omega = [0,1]^3$, the unit cube.

In each case, Ω is a set of points. We think of it as a **region**, and a random outcome is a **random point in that region**. **Region = population of points** (uncountably many of them), and every point is equally likely. This is the continuous analogue of our finite population: instead of 100 people, we have infinitely many points, but the logic of "favorable divided by total" still applies.

1.3.1 Probability as Measure

Example 1.3.1 (Estimating π)**.** Let X and Y be independent uniform random numbers in $[0,1]$. The point (X,Y) is a random point in $\Omega = [0,1]^2$. Let $A = \{(x,y) : x^2 + y^2 \leq 1\}$ be the quarter-circle of radius 1 centered at the origin.

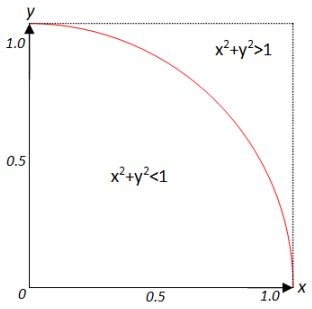


Figure 1.9: The quarter-circle A inside the unit square Ω.

By the same logic as before — favorable divided by total — the probability of landing in A is:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{area of } A}{\text{area of } \Omega} = \frac{\pi/4}{1} = \frac{\pi}{4}.$$

Here $|A|$ denotes the **size** (measure) of A — in two dimensions, this is the area. In one dimension it would be length, in three dimensions volume, and so on. The formula $P(A) = |A|/|\Omega|$ works in every dimension; we just need to use the right notion of "size."

The area of the quarter-circle is $\pi r^2/4 = \pi/4$ (since $r = 1$), and the area of the unit square is 1. So the probability of the random point falling inside the quarter-circle is exactly $\pi/4$.

In random variable notation, we can write:

$$P(X^2 + Y^2 \leq 1) = \pi/4, \qquad P(X^2 + Y^2 = 1) = 0.$$

We use capital letters for random variables. Note that a single curve has zero area, so hitting it exactly has probability zero. This is a distinctive feature of continuous probability: individual points have probability zero, and only regions with positive area (or length, or volume) have positive probability.

1.3.2 Conditional Probability in the Region Setting

The conditional probability formula works exactly the same way in the continuous setting. We just replace "counting" with "measuring area."

Within the region $\Omega = [0, 1]^2$, define two events:

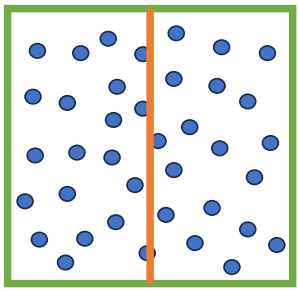


Figure 1.10: Event $A = \{(x, y) : x < 1/2\}$: the left half of the square.

$A = \{(x, y) : x < 1/2\}$, so $P(A) = P(X < 1/2) = |A|/|\Omega| = 1/2$. The left half of the square has area $1/2$, and the whole square has area 1.

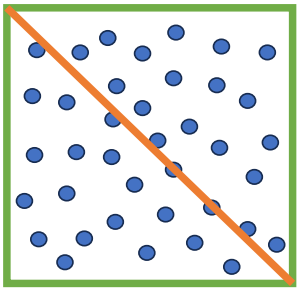


Figure 1.11: Event $B = \{(x, y) : x + y < 1\}$: the lower-left triangle.

$B = \{(x, y) : x + y < 1\}$, so $P(B) = P(X + Y < 1) = |B|/|\Omega| = 1/2$. This is the triangle below the line $x + y = 1$. Its area is $\frac{1}{2} \times 1 \times 1 = 1/2$.

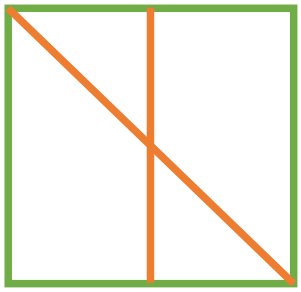


Figure 1.12: The intersection $A \cap B$ and the conditional probability.

Now, the intersection $A \cap B$ is the region where both $x < 1/2$ and $x + y < 1$. To find its area, we can think of the triangle B (area $1/2$) and subtract the small triangle in the right half (which has vertices at $(1/2, 0)$, $(1, 0)$, and $(1/2, 1/2)$, with area $1/8$). So $|A \cap B| = 1/2 - 1/8 = 3/8$. The conditional probability is:

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{3/8}{1/2} = \frac{3}{4}.$$

In random variable notation: $P(X < 1/2 \mid X + Y < 1) = 3/4$.

This can be interpreted in two ways:

1. Randomly throw a point into B (as if B is the sample space). What is the probability the point falls into A?
2. Throw many points into Ω. Among all the points that land in B, what fraction also belongs to A?

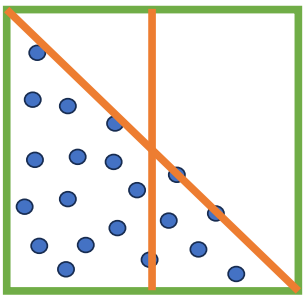


Figure 1.13: Among points falling in B, what fraction also falls in A?

Both interpretations give the same answer: $3/4$. The second interpretation connects conditional probability to long-run frequency, which we will explore in the next section.

1.3.3 Measuring by Counting

How do we compute the area of an irregular region? We **discretize**: divide the plane into tiny squares, and count how many of them fall inside the region. This is where continuous probability connects back to our original theme of counting.

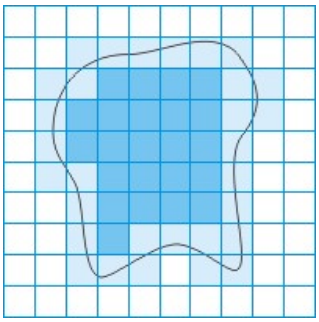


Figure 1.14: Inner and outer measure by discretization.

We create a finite population of tiny squares. The **area** equals the number of tiny squares times the area of each tiny square. As the squares get smaller, our approximation gets better.

Inner measure: fill the inside with tiny squares → gives a lower bound on the area.

Outer measure: cover the region from outside with tiny squares → gives an upper bound on the area.

Measurable: the region is measurable if the inner measure equals the outer measure as the squares become infinitesimally small. In other words, a region is measurable if we can pin down its area exactly by this procedure.

The collection of all measurable sets forms what is called a σ-algebra. The **integral** is the area under a curve, computed by this same limiting procedure. If you have studied Riemann integration in calculus, this is the same idea — we approximate the area under a curve by summing up the areas of thin rectangles, and the integral is the limit as the rectangles get infinitely thin.

The upshot is that "measuring" in the continuous case is ultimately a refined form of "counting" — we count tiny squares and take a limit. So the slogan "as long as you can count" extends seamlessly to the continuous setting.

1.3.4 Axioms of Probability

We can now state the axioms of probability. Probability is a **measure** — a way to assign sizes (counts, lengths, areas, volumes) to sets.

Axiom 0: $P(A) = |A|/|\Omega|$ in the equally likely scenario. This is our starting point.

Axiom 1: $P(\Omega) = 1$ (the total probability is 1). Something must happen — the outcome must lie somewhere in the sample space.

Axiom 2: $P(A) \geq 0$ for all events A (probabilities are non-negative). Probabilities cannot be negative.

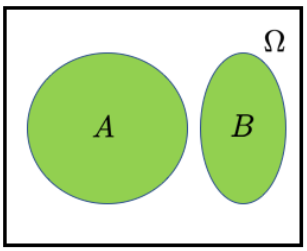


Figure 1.15: Two disjoint events: $P(A \cup B) = P(A) + P(B)$.

Axiom 3 (Additivity): If $A \cap B = \emptyset$ (empty), then $P(A \cup B) = P(A) + P(B)$. If two events cannot happen simultaneously (they are *disjoint* or *mutually exclusive*), then the probability of either one happening is the sum of their individual probabilities. This is intuitive: if 30 people are male-and-tall, and 10 people are female-and-tall, and no person is both male-and-tall and female-and-tall, then 40 people are tall.

Axiom 4: $P(A \mid B) = P(A \cap B)/P(B)$, assuming $P(B) > 0$. This is the definition of conditional probability.

Axiom 0 is the foundation; Axioms 1–3 are the standard Kolmogorov axioms (named after the Russian mathematician Andrei Kolmogorov, who formalized probability theory in the 1930s); Axiom 4 is the definition of conditional probability. All the properties of probability can be derived from these. Everything else in this book is a consequence.

1.4 Counting Repetitions and the Frequency Interpretation

We now come to one of the most important ideas in the course: the connection between probability and frequency through **repeated experiments**. This connection is what makes probability useful in the real world — it tells us that probabilities predict what actually happens when we repeat an experiment many times.

1.4.1 Long-Run Frequency

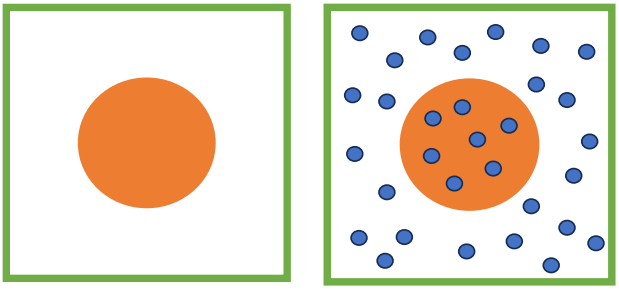


Figure 1.16: Left: a single random point in Ω. Right: many points thrown into Ω.

Throw n points into Ω. Let m be the number of points that fall into the event A. Then:

$$P(A) = \frac{|A|}{|\Omega|} \approx \frac{m}{n}.$$

As $n \to \infty$, $m/n \to P(A)$ in probability. This means that $P(A)$ can be interpreted as the **long-run frequency**: how often event A happens in the long run.

Here is the intuition. If you throw one point into Ω, you cannot predict where it lands. But if you throw a million points, about $P(A) \times 1{,}000{,}000$ of them will land in A, about $P(A) \times 1{,}000{,}000$ of them will land in A. The fraction m/n fluctuates from one experiment to the next, but it hovers close to $P(A)$, and the fluctuations get smaller as n increases.

1.4.2 Monte Carlo Estimation of π

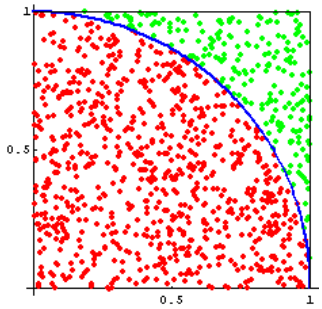


Figure 1.17: Monte Carlo estimation of π: throwing random points into Ω.

Returning to Example 1.3.1, throw n points into $\Omega = [0,1]^2$, and let m of them fall inside the quarter-circle A. Then:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\pi}{4} \approx \frac{m}{n}, \qquad \text{so} \qquad \hat{\pi} = \frac{4m}{n}.$$

This is the **Monte Carlo method**: estimating a quantity by random simulation. As $n \to \infty$, our estimate converges to the true value. With $n = 10{,}000$ random points, you might get $\hat{\pi} \approx 3.14$. With $n = 1{,}000{,}000$, you might get $\hat{\pi} \approx 3.1416$.

The power of Monte Carlo lies in its ability to work in *any* dimension. A deterministic method requires visiting all points on a grid. In 2 dimensions with 10 points per axis, that is $10^2 = 100$ points. In 3 dimensions, $10^3 = 1000$ points. In 10000 dimensions? 10^{10000} points — completely infeasible. But Monte Carlo can simply sample 1000 random points in the hyper-cube and give a useful estimate, regardless of dimension. This is why Monte Carlo methods are essential in modern science and engineering.

Example 1.4.1 (Buffon's needle)**.** In 1901, Lazzarini reportedly threw a needle $n = 3408$ times onto a ruled surface and obtained $\hat{\pi} = 355/113$, which is suspiciously accurate.

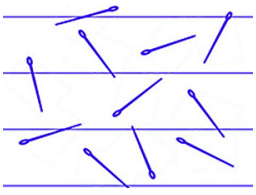


Figure 1.18: Buffon's needle experiment.

The result was too accurate! For fixed n, the count m is random. The ratio m/n fluctuates around $P(A)$. Only as $n \to \infty$ does $m/n \to P(A)$ in probability. This is the **law of large numbers**: $P(A)$ can be interpreted as long-run frequency — how often A happens in the long run. Lazzarini's suspiciously precise result suggests he may have stopped throwing once he hit the "right" answer, which is a form of cheating!

1.4.3 Fluctuations and the Population of Sequences

To understand *why* $m/n \to P(A)$, we use a powerful idea: the **population of sequences**. This is one of the most important conceptual tools in this course.

Repeat random sampling n times independently. Consider the sample space of all n-repetitions: this is the **hyper-population** of all possible sequences. Each sequence is a complete record of what happened in all n trials.

Among all these equally likely sequences, the overwhelming majority have m/n close to $P(A)$. Specifically:

- 99.999% of the sequences are "typical" — their frequency m/n is close to $P(A)$.
- Only a tiny fraction (say, 0.00000001%) are "atypical" — their frequencies are far from $P(A)$.

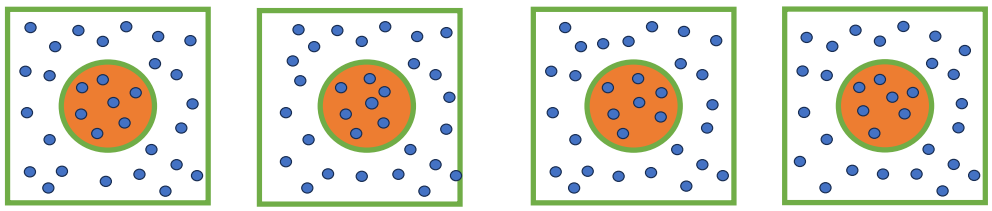


Figure 1.19: Typical sequences: m/n is close to $P(A)$.

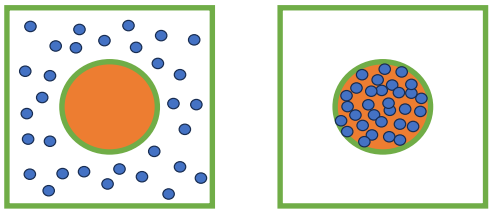


Figure 1.20: Atypical sequences: m/n is far from $P(A)$. These are very rare.

We can prove rigorously that:

$$P\left(\left|\frac{m}{n} - P(A)\right| > \epsilon\right) \to 0 \qquad \text{for any fixed } \epsilon > 0.$$

This is the **law of large numbers**. The fraction of representative sequences within the hyper-population approaches 1 as $n \to \infty$. In other words, if you pick a random sequence, it is overwhelmingly likely to be a typical one.

1.4.4 Conditional Probability via Repetitions

The frequency interpretation also applies to conditional probability. Consider throwing many points into Ω:

- How often does A happen? About $P(A)$ of the time.
- How often does B happen? About $P(B)$ of the time.

- When B happens, how often does A also happen? About $P(A \mid B)$ of the time.
- Among all the points that land in B, what is the fraction that belongs to A? About $P(A \mid B)$.

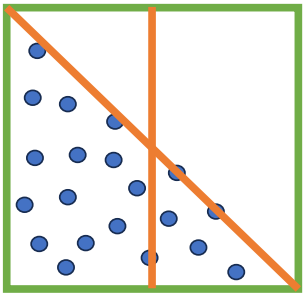


Figure 1.21: Counting repetitions for conditional probability in the region.

Note that regular probability is a special case of conditional probability: $P(A) = P(A \mid \Omega)$. The full sample space Ω is an implicit condition. When we write $P(A)$ without any condition, we are implicitly conditioning on the entire sample space — that is, assuming nothing is known.

With a fixed condition (within the same sub-population B), conditional probability behaves like regular probability. For example, $P(A^c) = 1 - P(A)$ becomes $P(A^c \mid B) = 1 - P(A \mid B)$. All the usual rules of probability hold within the conditional world, because conditioning on B simply means switching to a new sample space B.

1.5 Counting: Permutations and Combinations

To compute probabilities in the equally likely setting, we need to count. The formula $P(A) = |A|/|\Omega|$ requires us to count both $|A|$ and $|\Omega|$, and this can be tricky when the numbers are large. This section develops the basic counting tools: the multiplication principle, permutations, and combinations.

1.5.1 The Multiplication Principle

If experiment 1 has n_1 outcomes, and for each outcome of experiment 1, experiment 2 has n_2 outcomes, then the total number of all possible ordered pairs is $n_1 \times n_2$.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Figure 1.22: Rolling a die twice: $6 \times 6 = 36$ ordered pairs (table form).

For example, if you roll a die twice, the first roll has 6 outcomes and the second roll has 6 outcomes, giving $6 \times 6 = 36$ possible ordered pairs. The pair $(3, 5)$ is different from $(5, 3)$ because the order matters — the first number is the first roll and the second number is the second roll.

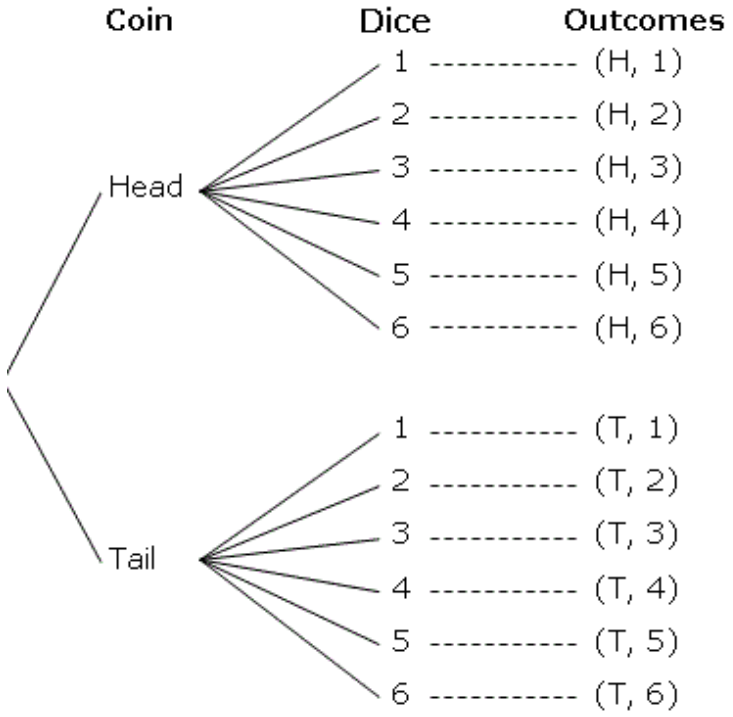


Figure 1.23: Tree diagram: flip a coin and roll a die.

A tree diagram makes this visual. Each branch of the tree represents one possible outcome. The total number of paths through the tree — from root to leaf — equals the total number of outcomes.

This extends to ordered sequences: if we have experiments with $n_1, n_2, \ldots, n_k$ outcomes respectively, the total number of ordered k-tuples is $n_1 \times n_2 \times \cdots \times n_k$.

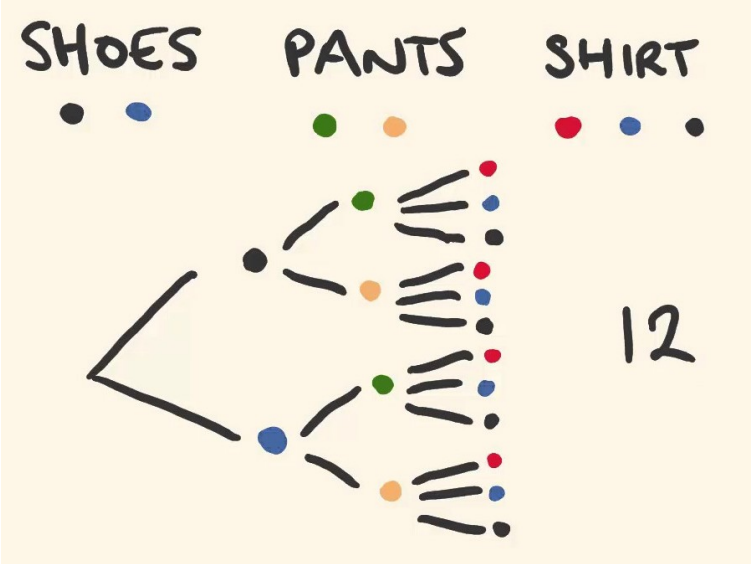


Figure 1.24: Tree diagram: ordered triplet.

1.5.2 Permutations

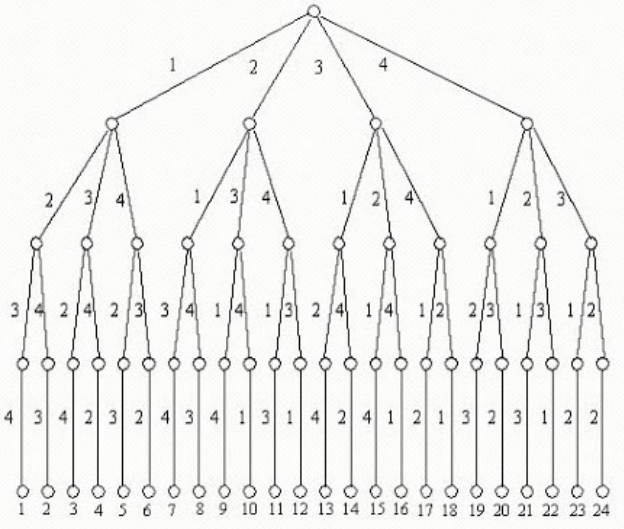


Figure 1.25: Permutations: choosing k cards from n, order matters.

Given n different cards, choose k of them. Order matters. How many different ordered sequences are possible? For the first card, we have n choices. For the second card, we have $n-1$ choices (one card has been removed). For the third, $n-2$ choices, and so on. By the multiplication principle:

$$P_{n,k} = n(n-1)(n-2)\cdots(n-k+1).$$

For example, $P_{4,2} = 4 \times 3 = 12$. From 4 cards, there are 12 ways to choose an ordered pair. The number of ways to permute all n objects is:

$$P_{n,n} = n!$$

This counts how many different ways there are to permute (rearrange) things. For instance, $4! = 24$ means there are 24 different ways to arrange 4 cards in a row.

1.5.3 Combinations

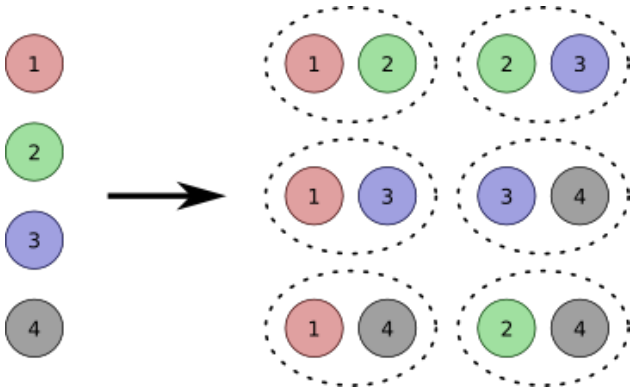


Figure 1.26: Combinations: choosing k balls from n, order does NOT matter.

If order does *not* matter, the number of ways to choose k objects from n is the **binomial coefficient**:

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

For example, $\binom{4}{2} = \frac{4\times 3}{2} = 6$.

Figure 1.27: Each combination corresponds to $k!$ permutations.

The division by $k!$ accounts for the fact that each combination of k objects corresponds to $k!$ different permutations (orderings). For example, the combination $\{A, B\}$ corresponds to two permutations: (A, B) and (B, A). Since we do not care about order, we divide by $2! = 2$ to avoid counting $\{A, B\}$ twice. More generally, each unordered group of k objects can be arranged in $k!$ ways, so we divide the number of permutations by $k!$ to get the number of combinations.

1.6 Coin Flipping and the Binomial Distribution

Coin flipping is the simplest non-trivial probability experiment, and it leads to one of the most important distributions in probability: the binomial distribution. Every concept in this section will reappear throughout the book.

Example 1.6.1 (Coin flipping)**.** Consider flipping a fair coin.

(4.1) Flip a coin $\to$ head or tail $\to$ 1 or 0.

(4.2) Flip a coin twice $\to$ (head, head), (head, tail), (tail, head), (tail, tail) $\to$ 11, 10, 01, 00.

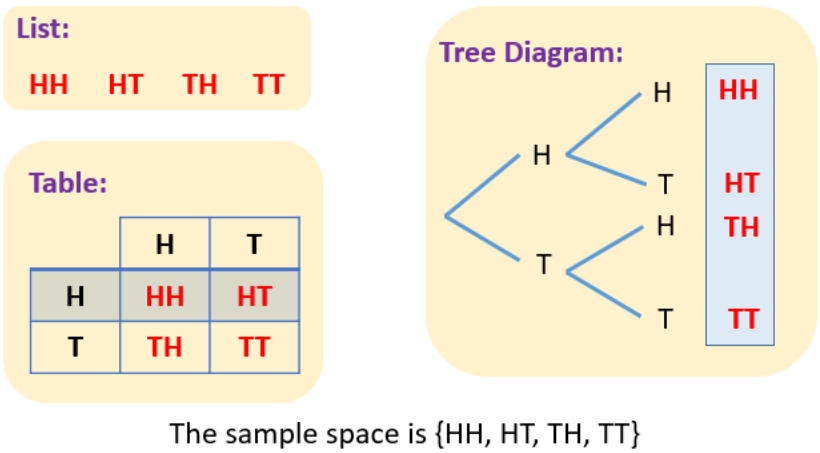


Figure 1.28: Sample spaces for 1 and 2 coin flips.

1.6.1 The Sample Space of Sequences

(4.3) Flip a fair coin n times $\to 2^n$ binary sequences.

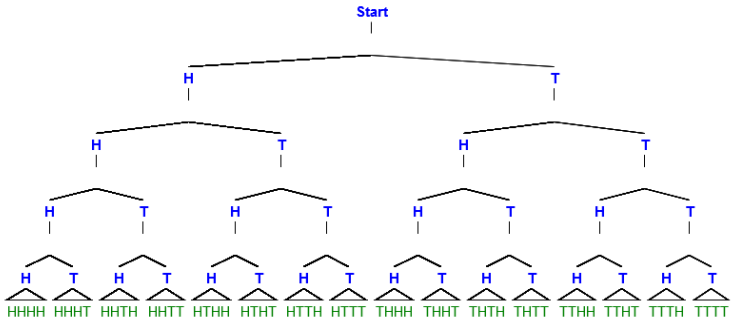


Figure 1.29: The tree of all binary sequences from n coin flips.

The sample space Ω consists of all 2^n binary sequences. Each $\omega \in \Omega$ is a specific sequence of heads and tails. For $n = 4$, there are $2^4 = 16$ sequences, ranging from HHHH to TTTT.

We define $Z_i(\omega) = 1$ if the i-th flip is heads, and $Z_i(\omega) = 0$ if it is tails. The fundamental observation is: **we randomly pick one sequence from the 2^n equally likely sequences**. Each sequence has probability $1/2^n$.

1.6.2 Counting Sequences with k Heads

Let $X(\omega)$ be the total number of heads in the sequence ω:

$$X(\omega) = Z_1(\omega) + Z_2(\omega) + Z_3(\omega) + Z_4(\omega).$$

This is our first example of a random variable that is a *sum* of simpler random variables. The value of X tells us how many 1's appear in the sequence, without caring about their positions.

Let $A_k = \{\omega : X(\omega) = k\}$ be the event that there are exactly k heads. Then:

$$P(A_k) = P(\{\omega : X(\omega) = k\}) = P(X = k) = p_k.$$

HHHH, THHH, HTHT, TTHT,
HHHT, HHTT, THHT, THTT,
HHTH, TTHH, HTTH, HTTT,
HTHH, THTH, TTTH, TTTT

Figure 1.30: The 16 sequences from 4 coin flips, organized by number of heads.

For $n = 4$: $(p_k, k = 0, 1, 2, 3, 4) = (1, 4, 6, 4, 1)/16$.

Let us verify this for $k = 2$. How do we get $|A_2| = 6$? We compute: $|A_2| = \binom{4}{2} = \frac{4\times 3}{2} = 6$. The logic: we have 4 positions in the sequence. Choose 2 of them to be heads, and the rest are automatically tails. The number of ways to choose 2 positions from 4 is $\binom{4}{2} = 6$.

HHHH, THHH, HTHT, TTHT,
HHHT, HHTT, THHT, THTT,
HHTH, TTHH, HTTH, HTTT,
HTHH, THTH, TTTH, TTTT

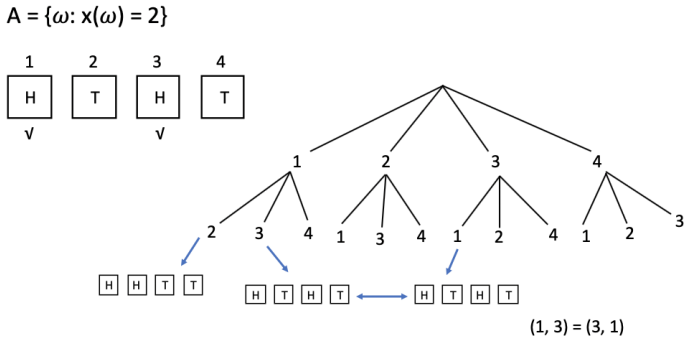


Figure 1.31: The 6 sequences with exactly 2 heads out of 4 flips, and the corresponding combinations.

Important: Do not confuse the order of picking the blank positions (which is irrelevant, since we use combinations) with the order of the coin flippings (which is fixed by the sequence). We use combinations because we are choosing *which positions* get heads, not *in what order* we fill them.

In general, for flipping a fair coin n times independently:

$$P(A_k) = P(\{\omega : X(\omega) = k\}) = P(X = k) = \frac{\binom{n}{k}}{2^n}.$$

The numerator $\binom{n}{k}$ counts the number of sequences with exactly k heads, and the denominator 2^n is the total number of sequences.

1.6.3 The Population of Sequences

The concept of the **population of sequences** is central to this course. When we flip a fair coin n times independently, the sample space is:

$$\Omega_n = \Omega_1 \times \Omega_1 \times \cdots \times \Omega_1 = \Omega_1^n,$$

the n-fold product of the base sample space $\Omega_1 = \{\text{head, tail}\}$. Each element of Ω_n is a complete sequence of n outcomes.

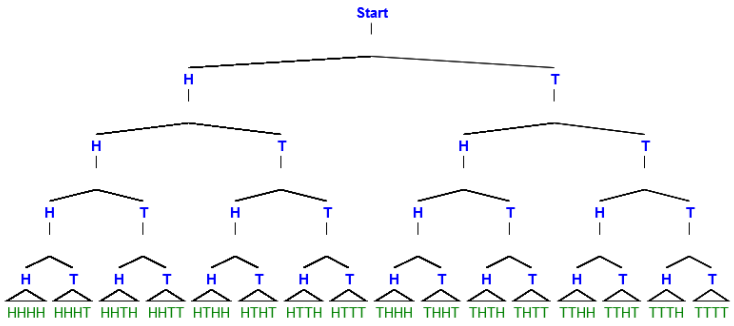


Figure 1.32: The hyper sample space Ω_n of coin flip sequences.

The number of elements is $|\Omega_n| = 2^n$. The key principle is:

Equally likely outcomes in Ω_1 + independent repetitions = equally likely sequences in Ω_n.

Why is this true? Because each flip is independent and has the same probabilities, every sequence has the same probability: $(1/2)^n$. The sequence HHHT is just as likely as HTHT or TTTT — each has probability $1/2^n$. This principle is universal. It applies to any base experiment:

Die rolling: $\Omega_1 = \{1, 2, \ldots, 6\}$. Roll n times independently. $|\Omega_n| = 6^n$. The hyper sample space is the population of all 6^n sequences.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Figure 1.33: Die rolling: $|\Omega_1| = 6$, giving $|\Omega_n| = 6^n$ sequences.

Population sampling: Randomly sample a person from a population of N (e.g., 300 million) people. Repeat n (e.g., 1000) times independently.

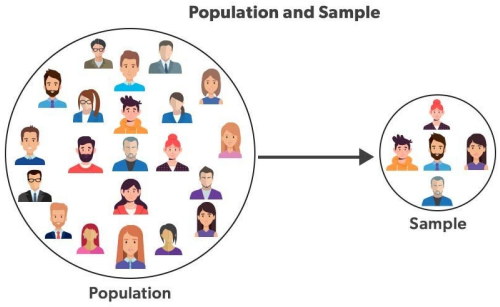


Figure 1.34: Sampling from a population: $|\Omega_1| = N$, giving $|\Omega_n| = N^n$ sequences.

$|\Omega_n| = N^n$ (e.g., 300m^{1000}). The hyper sample space Ω_n is the hyper-population of sequences. This is an astronomically large number, but the logic is the same: every sequence is equally likely.

Region sampling: Randomly sample a point from a region. Repeat n times independently.

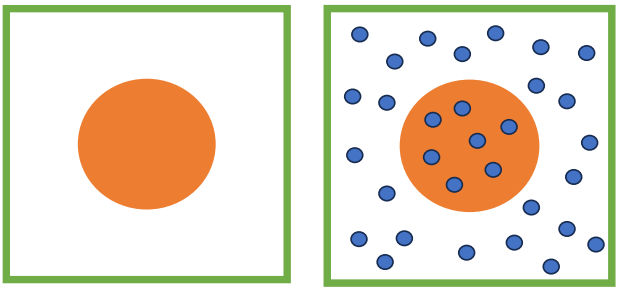


Figure 1.35: Region sampling: $\Omega_1 = [0,1]^2$, giving $\Omega_n = [0,1]^{2n}$.

$\Omega_n = \Omega_1^n$. The point $(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ is a point in Ω_n.
For the population setting, each sequence $\omega = (a_1, a_2, \ldots, a_n)$ has:

$$P(\omega) = P(a_1)P(a_2)\cdots P(a_n) = \frac{1}{N} \times \frac{1}{N} \times \cdots \times \frac{1}{N} = \frac{1}{N^n}.$$

This factorization is the mathematical expression of independence: the probability of a sequence is the product of the probabilities of its individual elements.

1.6.4 Survey Sampling and the Binomial Distribution

Consider a population of N people, of whom M are male. We sample n people independently (with replacement). Let $p = M/N$ be the probability of selecting a male on any single draw.

For a sequence ω, let $X(\omega)$ be the number of males. Let $A_m = \{\omega : X(\omega) = m\}$ be the set of sequences with m males. How many sequences have exactly m males?

We need n blanks (one for each draw). Choose m of them for males (the rest $n - m$ blanks are for females). Each male blank has M choices (any of the M males in the population). Each female blank has $N - M$ choices (any of the $N - M$ females). By the multiplication principle:

$$|A_m| = \binom{n}{m} M^m (N - M)^{n-m}.$$

The probability is:

$$\begin{aligned} P(X = m) = \frac{|A_m|}{|\Omega_n|} &= \frac{\binom{n}{m} M^m (N-M)^{n-m}}{N^n} \\ &= \binom{n}{m} \left(\frac{M}{N}\right)^m \left(\frac{N-M}{N}\right)^{n-m} = \binom{n}{m} p^m (1-p)^{n-m}. \end{aligned}$$

This is the **binomial distribution**. It applies to coin flipping ($p = 1/2$), survey sampling ($p = M/N$), and Monte Carlo ($p = \pi/4$). The binomial distribution is universal — it describes the count of "successes" in any sequence of independent trials with the same probability of success.

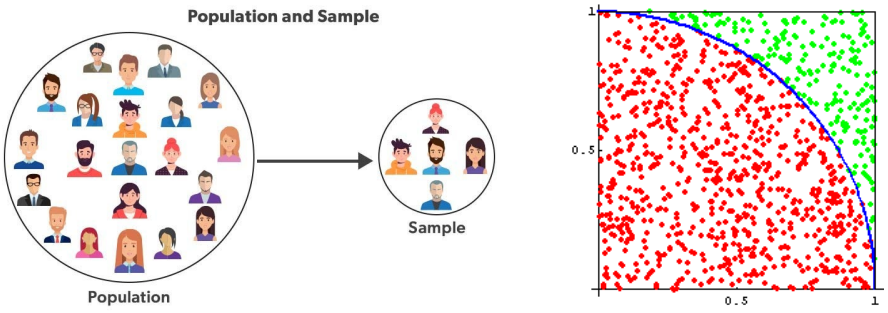


Figure 1.36: Left: survey sampling. Right: Monte Carlo for π.

1.6.5 Law of Large Numbers via Counting

Among all N^n sequences in the hyper-population Ω_n, define the set of **representative sequences**:

$$A = \left\{ \omega : \left| \frac{X(\omega)}{n} - p \right| \leq 0.01 \right\}.$$

These are the sequences whose frequency of males is within 1% of the true probability p. The law of large numbers says:

$$P(A) = \frac{|A|}{|\Omega_n|} = \sum_{x \in [n(p-0.01),\, n(p+0.01)]} P(X = x) \to 1 \quad \text{as } n \to \infty.$$

For example, with $n = 100$, we need $x \in [49, 51]$; with $n = 1000$, $x \in [490, 510]$.

In words: the proportion of representative sequences within the hyper-population approaches 1. As we take more samples, it becomes overwhelmingly likely that the observed frequency is close to the true probability.

$X/n \to p$ in probability: the frequency converges to the probability.

Remark 1.6.1 (The correct definition of probability)**. Right:** Probability is defined as population proportion, $P(A) = |A|/|\Omega|$, or as a normalized measure, or as subjective belief or common sense of uncertainty.

Wrong: Probability is defined as long-run frequency, $P(A) = \lim_{n\to\infty} X/n$, under independent repetitions of the same experiment.

The frequency definition is problematic because: (1) the limit does not always exist, nor is it the same for every sequence of repetitions; (2) the notion of independence is not defined within the frequency framework.

Right: Start with the **hyper-population of sequences of repetitions**. Under uniformity (Axiom 0) and independence, all sequences are equally likely. Then prove that the proportion of representative sequences within this hyper-population approaches 1 as $n \to \infty$. This is a *theorem* (the law of large numbers), not a definition.

1.6.6 Special Case: Fair Coin

For $p = 1/2$ (or equivalently $N = 2$):

$$p(x) = P(X = x) = \frac{\binom{n}{x}}{2^n}, \qquad x = 0, 1, \ldots, n.$$

HHHH, THHH, HTHT, TTHT,
HHHT, HHTT, THHT, THTT,
HHTH, TTHH, HTTH, HTTT,
HTHH, THTH, TTTH, TTTT

Figure 1.37: Sequences from fair coin flips.

Among all 2^n sequences, let

$$A = \left\{ \omega : \left| \frac{X(\omega)}{n} - \frac{1}{2} \right| \leq 0.01 \right\}$$

consist of representative sequences. Then:

$$P(A) = \sum_{x \in [n(1/2-0.01),\, n(1/2+0.01)]} \frac{\binom{n}{x}}{2^n} \to 1.$$

For $n = 100$: $x \in [49, 51]$. For $n = 1000$: $x \in [490, 510]$. The frequency $X/n \to 1/2$ in probability.

1.6.7 Random Walk

A particularly beautiful application of coin flipping is the **random walk**. Imagine standing at position 0 on a number line. At each step, you flip a fair coin: if heads, move one step to the right $(+1)$; if tails, move one step to the left (-1).

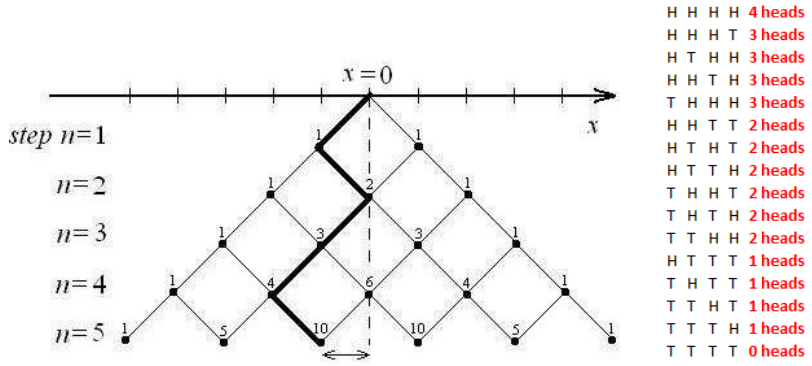


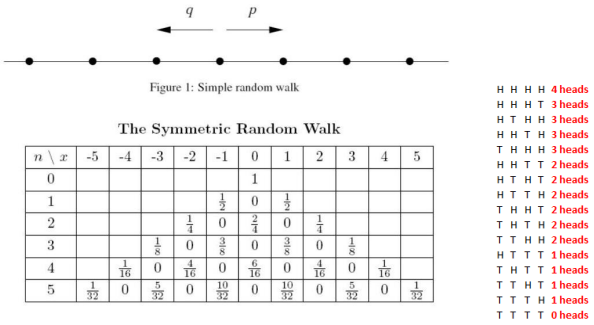
Figure 1.38: A random walk: the position after n steps depends on the number of heads.

Walk n steps. If the number of heads is $X = x$, then you moved right x times and left $(n - x)$ times. Your final position is:

$$Y = x - (n - x) = 2x - n, \qquad \text{i.e.,} \qquad x = (Y + n)/2.$$

The probability of ending at position y is:

$$p_Y(y) = P(Y = y) = P(X = x) = p_X(x) = \frac{\binom{n}{x}}{2^n} = \frac{\binom{n}{(y+n)/2}}{2^n}.$$



$n \setminus x$	-5	-4	-3	-2	-1	0	1	2	3	4	5
0						1					
1					$\frac{1}{2}$	0	$\frac{1}{2}$				
2				$\frac{1}{4}$	0	$\frac{2}{4}$	0	$\frac{1}{4}$			
3			$\frac{1}{8}$	0	$\frac{3}{8}$	0	$\frac{3}{8}$	0	$\frac{1}{8}$		
4		$\frac{1}{16}$	0	$\frac{4}{16}$	0	$\frac{6}{16}$	0	$\frac{4}{16}$	0	$\frac{1}{16}$	
5	$\frac{1}{32}$	0	$\frac{5}{32}$	0	$\frac{10}{32}$	0	$\frac{10}{32}$	0	$\frac{5}{32}$	0	$\frac{1}{32}$

Figure 1.39: Multiple random walk trajectories.

The random walk is a fundamental model in probability, physics, and finance. Stock prices, the position of a diffusing particle, and the trajectory of a drunk person walking down a street can all be modeled as random walks. We will return to this model when we study Brownian motion in Chapter 4.

1.6.8 Pascal's Triangle and the Galton Board

The binomial coefficients are organized in **Pascal's triangle**, where each entry is the sum of the two entries above it: $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.

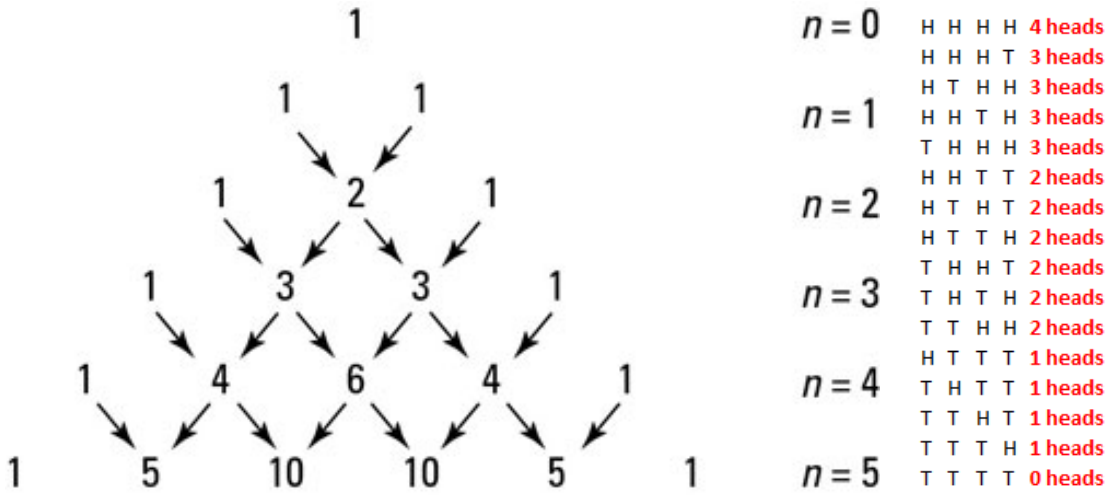


Figure 1.40: Pascal's triangle: the n-th row gives $\binom{n}{k}$ for $k = 0, 1, \ldots, n$.

Why does this recursion hold? Think of it this way: a sequence of n coin flips with exactly k heads either starts with a head (and the remaining $n-1$ flips have $k-1$ heads, which happens $\binom{n-1}{k-1}$ ways) or starts with a tail (and the remaining $n-1$ flips have k heads, which happens $\binom{n-1}{k}$ ways).

The **Galton board** (named after Sir Francis Galton) is a physical device that realizes the random walk. A ball is dropped at the top and bounces left or right at each peg, like a coin flip at each step.

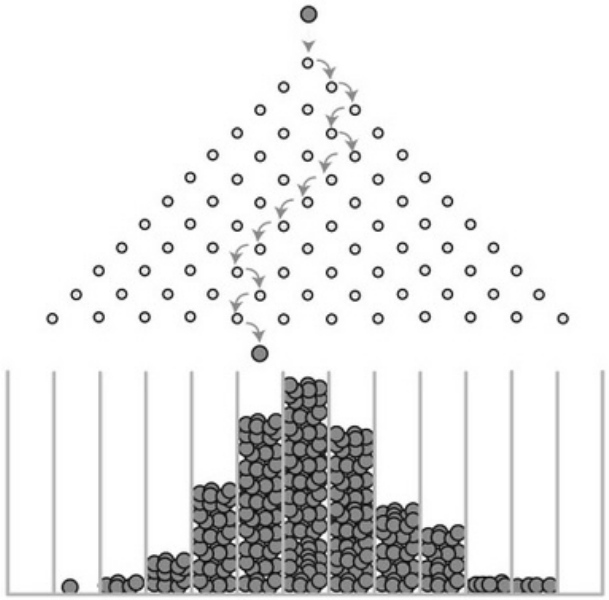


Figure 1.41: The Galton board: balls follow random paths and accumulate in a bell-shaped histogram.

All 2^n paths are equally likely — they form a **population of trajectories**. The number of paths that end up in the x-th bin is $\binom{n}{x}$. X is the destination, and $p(x) = P(X = x) = \binom{n}{x}/2^n$.

If we drop 1 million balls, how often do balls fall into the x-th bin? The histogram approaches the binomial distribution, and as n grows, the bell-shaped curve of the normal distribution emerges. This is one of the earliest demonstrations of the central limit theorem — one of the crown jewels of probability theory, which we will prove in Chapter 3.

1.7 Markov Chains

Our next topic is a beautiful generalization of the random walk: the **Markov chain**. In a random walk, the position at the next step depends only on the current position and the coin flip — it does not depend on how you got to the current position. Markov chains generalize this idea to systems that can be in one of several states.

1.7.1 Random Walk and Transition Probability

Recall the random walk: either go forward or backward by flipping a fair coin.

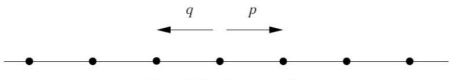


Figure 1: Simple random walk

The Symmetric Random Walk

$n \setminus x$	-5	-4	-3	-2	-1	0	1	2	3	4	5
0						1					
1					$\frac{1}{2}$	0	$\frac{1}{2}$				
2				$\frac{1}{4}$	0	$\frac{2}{4}$	0	$\frac{1}{4}$			
3			$\frac{1}{8}$	0	$\frac{3}{8}$	0	$\frac{3}{8}$	0	$\frac{1}{8}$		
4		$\frac{1}{16}$	0	$\frac{4}{16}$	0	$\frac{6}{16}$	0	$\frac{4}{16}$	0	$\frac{1}{16}$	
5	$\frac{1}{32}$	0	$\frac{5}{32}$	0	$\frac{10}{32}$	0	$\frac{10}{32}$	0	$\frac{5}{32}$	0	$\frac{1}{32}$

H H H H 4 heads
H H H T 3 heads
H T H H 3 heads
H H T H 3 heads
T H H H 3 heads
H H T T 2 heads
H T H T 2 heads
H T T H 2 heads
T H H T 2 heads
T H T H 2 heads
T T H H 2 heads
H T T T 1 heads
T H T T 1 heads
T T H T 1 heads
T T T H 1 heads
T T T T 0 heads

Figure 1.42: Random walk: $X_t = Z_1 + Z_2 + \cdots + Z_t$.

The position at time t is $X_t = Z_1 + Z_2 + \cdots + Z_t$, where $Z_k = +1$ or -1 with probability $1/2$ each. The update rule is:

$$X_{t+1} = X_t + Z_{t+1}.$$

The **transition probabilities** describe how the system moves from one state to the next:

$$P(X_{t+1} = x + 1 \mid X_t = x) = P(X_{t+1} = x - 1 \mid X_t = x) = 1/2.$$

Notice that these probabilities depend only on the current position x, not on how we got there. This is the **Markov property**: the future depends on the present but not on the past.

Example 1.7.1 (Random walk over three states)**.** Consider a particle that can be in one of three states $\{1, 2, 3\}$. With probability $1/2$, it stays in its current state. With probability $1/4$, it goes to each of the other two states.

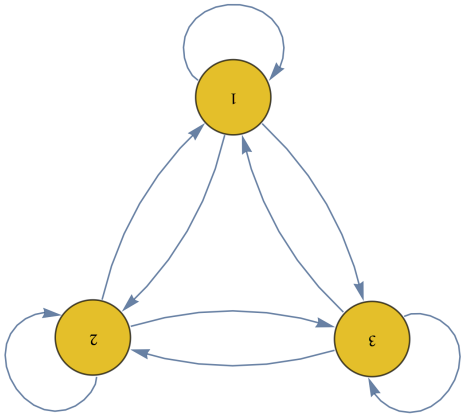


Figure 1.43: Random walk over three states with transition probabilities.

The transition probabilities are:

$$K_{ij} = P(X_{t+1} = j \mid X_t = i).$$

This is a **forward conditional probability**, from cause (current state) to effect (next state). The **Markov property** says: past history before X_t does not matter. Only the present state matters for predicting the future.

1.7.2 Population Migration Interpretation

To build intuition for Markov chains, imagine a vivid picture: 1 million people migrating among three cities. At each time step, for each state, half of the people stay, and a quarter go to each of the other two states. This gives a vivid picture: we are tracking the evolution of a **population** of people across states. We can think of 1 million trajectories unfolding simultaneously.

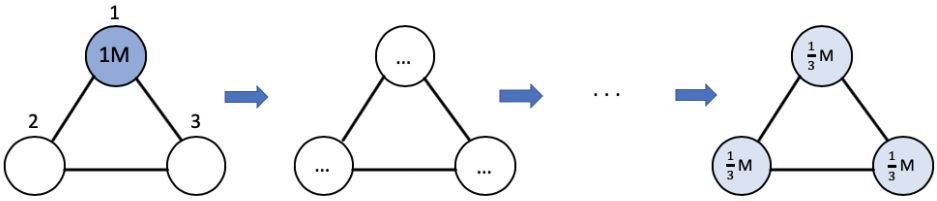


Figure 1.44: Population migration among three states.

This population migration picture is extremely useful. Instead of thinking about a single particle jumping randomly between states (which is hard to visualize), think about millions of people flowing between cities according to fixed proportions. The proportions are deterministic even though each individual's trajectory is random.

1.7.3 Transition Matrix

The transition probabilities are organized into a **transition matrix**:

$$\mathbf{K} = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}.$$

Each row sums to 1 (the particle must go somewhere). The entry K_{ij} in row i, column j is the probability of going from state i to state j. Such a matrix, where all entries are non-negative and each row sums to 1, is called a **stochastic matrix**.

1.7.4 Marginal Probability and Total Probability

Let $p_i^{(t)} = P(X_t = i)$ be the probability of being in state i at time t. In the population migration picture, $p_i^{(t)}$ is the number of people (in millions) in state i at time t. The vector $\mathbf{p}^{(t)} = (p_1^{(t)}, p_2^{(t)}, p_3^{(t)})$ describes the distribution at time t.

How does the distribution evolve? The number of people in state j at time $t+1$ equals the sum over all states i of the people in state i at time t who move to j:

$$p_j^{(t+1)} = \sum_i p_i^{(t)} K_{ij}.$$

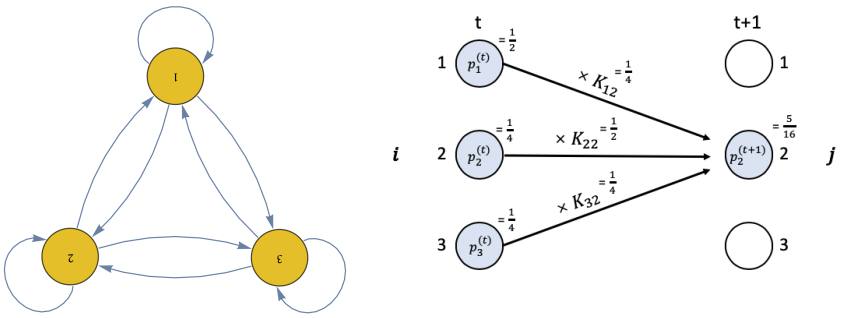


Figure 1.45: Population migration: updating the distribution.

This update rule is derived from two fundamental rules:

Chain rule: $P(X_{t+1} = j, X_t = i) = P(X_t = i)\,P(X_{t+1} = j \mid X_t = i) = p_i^{(t)} K_{ij}$. The probability of being in state i *and then* moving to state j is the probability of being in i times the probability of the transition.

Rule of total probability: $P(X_{t+1} = j) = \sum_i P(X_{t+1} = j, X_t = i)$. To find the total probability of being in state j at time $t+1$, we add up all the ways to get there — from state 1, from state 2, and from state 3.

We add up the probabilities of all the alternative chains of events that lead to state j.

1.7.5 Stationary Distribution

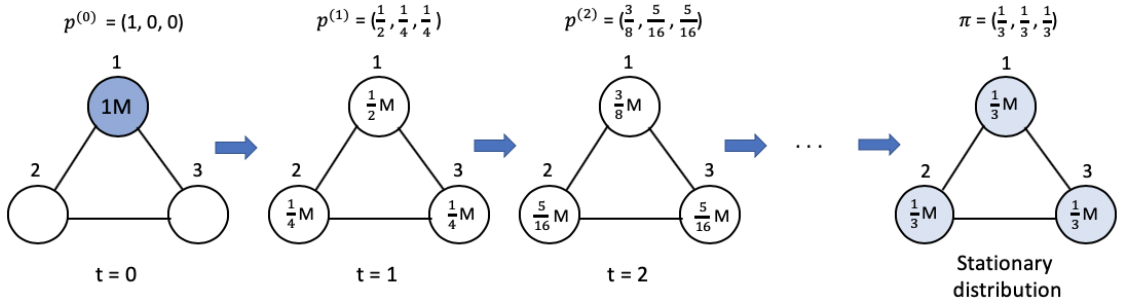


Figure 1.46: The distribution $\mathbf{p}^{(t)}$ converges to the stationary distribution $\boldsymbol{\pi}$.

As $t \to \infty$, a remarkable thing happens: the distribution converges. No matter where the particle (or the population) starts, after enough time, the distribution settles into a fixed pattern. This limiting distribution $\boldsymbol{\pi}$ is the **stationary distribution**, satisfying:

$$\pi_j = \sum_i \pi_i K_{ij}, \qquad \text{i.e.,} \qquad \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{K}.$$

This equation says: the stationary distribution is unchanged by one step of the Markov chain. The flow of people into each state exactly balances the flow out. No matter where the particle starts, it eventually reaches the same long-run distribution. This is the "arrow of time" in Markov chains.

1.7.6 Matrix Multiplication and Eigen-Analysis

In matrix notation, the update rule is:

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)}\mathbf{K}, \qquad \text{so} \qquad \mathbf{p}^{(t)} = \mathbf{p}^{(0)}\mathbf{K}^t \to \boldsymbol{\pi}.$$

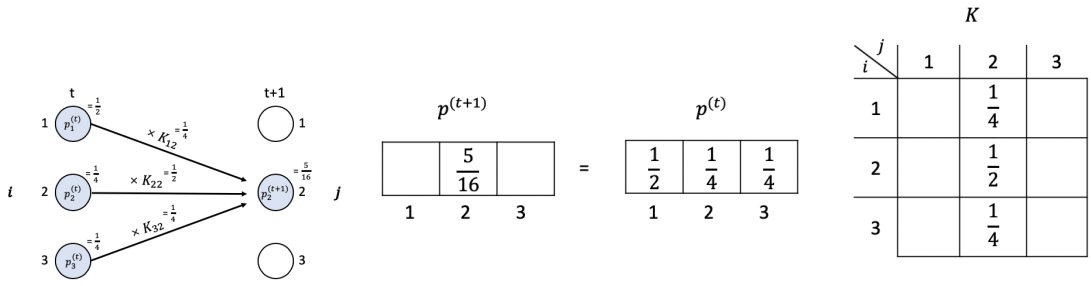


Figure 1.47: Matrix powers and convergence to the stationary distribution.

The convergence can be understood through **diagonalization**: $\mathbf{K} = \mathbf{PDP}^{-1}$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues. Then:

$$\mathbf{K}^t = \mathbf{PDP}^{-1}\,\mathbf{PDP}^{-1}\cdots\mathbf{PDP}^{-1} = \mathbf{PD}^t\mathbf{P}^{-1}.$$

The largest eigenvalue is 1, and $1^t = 1$. The second-largest eigenvalue is less than 1 (say, 0.99), and $0.99^t \to 0$. So the contributions from all eigenvalues except 1 vanish, and $\mathbf{p}^{(t)} \to \boldsymbol{\pi}$. The speed of convergence is controlled by the second-largest eigenvalue — the closer it is to 1, the slower the convergence.

1.7.7 Marginal, Conditional, and Joint Distributions

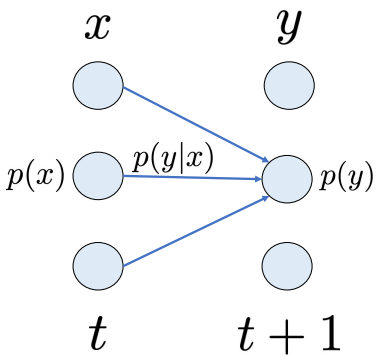


Figure 1.48: A Markov chain as a directed graphical model: $X_t \to X_{t+1}$.

Let us organize the key distributions for a Markov chain:

Marginal: $p_t(x) = P(X_t = x)$, $p_{t+1}(y) = P(X_{t+1} = y)$. These describe the distribution at a single time step.

Conditional (forward): $p(y \mid x) = P(X_{t+1} = y \mid X_t = x)$. Here x is the cause and y is the effect. The forward conditional goes from cause to effect — it is given or learned.

Joint: $p(x, y) = P(X_t = x, X_{t+1} = y)$. This describes the joint distribution of consecutive states.

Chain rule: $p(x, y) = p_t(x)\, p(y \mid x)$. The joint probability factors as the marginal times the conditional.

Rule of total probability: $p_{t+1}(y) = \sum_x p(x, y) = \sum_x p_t(x)\, p(y \mid x)$. To find the marginal at time $t + 1$, we sum over all possible states at time t.

We add up probabilities of alternative chains of events.

1.7.8 Google PageRank

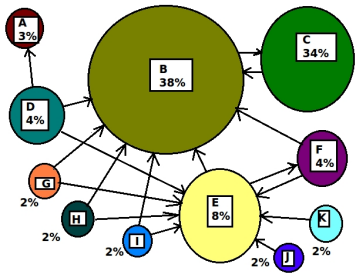


Figure 1.49: A web graph: pages linked to each other.

The same framework powers Google's PageRank algorithm, which revolutionized web search. Web pages are states, and hyperlinks define transition probabilities. A "random surfer" follows links randomly: at each page, they click a random link to move to a new page.

The stationary distribution π_i gives the proportion of time spent on page i in the long run — this is the page's "popularity." The popularity of page i depends on the popularities of all pages that link to i:

$$\pi_j = \sum_i \pi_i K_{ij}.$$

Pages with many incoming links from popular pages are themselves popular. This is the essence of PageRank: importance flows through the web graph, and the stationary distribution captures the equilibrium.

1.8 Conditional Reasoning and Bayes' Rule

We now come to one of the most important and frequently misunderstood topics in probability: **Bayes' rule**. This rule tells us how to reverse the direction of conditional probability — going from "cause given effect" to "effect given cause," or vice versa.

Example 1.8.1 (Rare disease)**.** 1% of the population has a rare disease. A random person goes through a test. If the person has the disease, there is a 90% chance the test is positive. If the person does not have the disease, there is a 90% chance the test is negative. If tested positive, what is the chance he or she has the disease?

Before reading the solution, take a moment to guess. Many people guess around 90%, reasoning that the test is "90% accurate." The true answer is much lower, and understanding why is one of the most valuable lessons in this course.

The given information is:

$$P(D) = 1\%, \quad P(N) = 99\%.$$
$$\textbf{Forward:} \quad P(+ \mid D) = 90\%, \quad P(- \mid N) = 90\%.$$

The question asks for the **backward** direction:

$$\textbf{Backward:} \quad P(D \mid +) = ?$$

1.8.1 Forward vs. Backward Conditioning

The given probabilities go from cause to effect (forward). The question asks to go from effect to cause (backward). This asymmetry is at the heart of Bayes' rule.

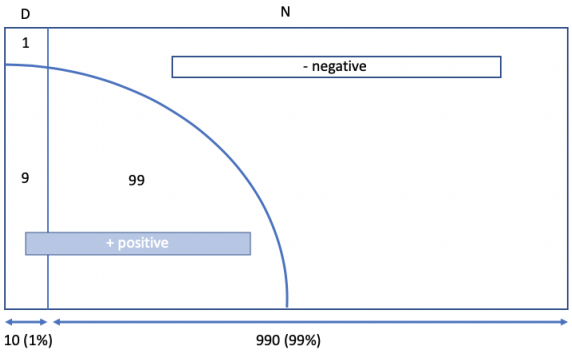


Figure 1.50: Tree diagram for the rare disease. Among 1000 people: 10 have disease, 990 don't. Of the 10 with disease, 9 test positive. Of the 990 without, 99 test positive.

Think of 1000 people. Of these, 10 have the disease (1%). Of the 10 diseased, 9 test positive (90%). Of the 990 healthy, 99 test positive (10% false positive rate). Among all $9 + 99 = 108$ who test positive, only 9 actually have the disease:

$$P(D \mid +) = \frac{9}{9 + 99} = \frac{1}{12} \approx 8.3\%.$$

This is much lower than the 90% accuracy of the test! The reason is that the disease is rare, so even a small false-positive rate produces many false positives. Among the 990 healthy people, 10% test positive, giving 99 false positives — far more than the 9 true positives.

This illustrates a general principle: $P(\text{alarm} \mid \text{fire})$ and $P(\text{fire} \mid \text{alarm})$ are very different quantities. A fire alarm might go off 95% of the time when there is a fire, but if most alarm sounds are caused by burnt toast, the probability that there is actually a fire when the alarm sounds might be quite low.

1.8.2 Chain Rule, Rule of Total Probability, and Bayes' Rule

The computation above uses three fundamental rules:

Chain rule: The probability of two events happening together equals the probability of the first times the conditional probability of the second given the first.

$$P(D \cap +) = P(D)\, P(+ \mid D) = 1\% \times 90\%.$$
$$P(N \cap +) = P(N)\, P(+ \mid N) = 99\% \times 10\%.$$

Rule of total probability: To find the overall probability of testing positive, we add up all the ways it can happen — either through having the disease or through not having it.

$$P(+) = P(D \cap +) + P(N \cap +) = 1\% \times 90\% + 99\% \times 10\%.$$

Bayes' rule: Now we can reverse the conditioning.

$$P(D \mid +) = \frac{P(D \cap +)}{P(+)} = \frac{9}{9+99} = \frac{1}{12}.$$

The chain rule says: population proportion of tall males = proportion of males × proportion of tall among males.

The rule of total probability says: add up the probabilities of alternative chains of events.

We can generalize the chain rule to chains of multiple events:

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

This factorization is the basis of many important models, including GPT and Bayes networks.

1.8.3 Random Variables and Probability Mass Functions

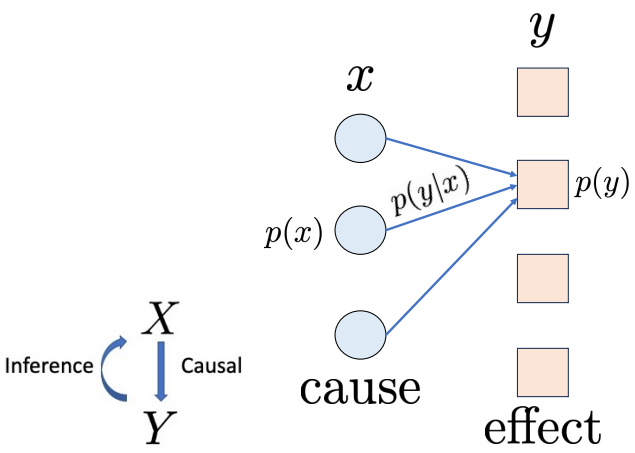


Figure 1.51: Cause X and effect Y: a directed graphical model.

In the language of random variables, let X be the cause and Y be the effect:

Marginal (prior): $p(x) = P(X = x)$, marginal $p(y) = P(Y = y)$.

Conditional: Forward generation $p(y \mid x) = P(Y = y \mid X = x)$.

Backward inference $p(x \mid y) = P(X = x \mid Y = y)$.

Chain rule: Joint $p(x, y) = p(x)\, p(y \mid x)$.

Rule of total probability: Marginal $p(y) = \sum_x p(x, y) = \sum_x p(x)\, p(y \mid x)$.

1.8.4 Bayes' Rule in Full Generality

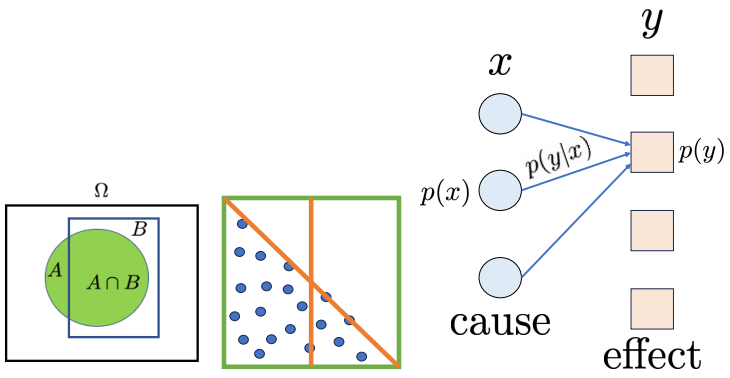


Figure 1.52: Bayes' rule connects forward and backward conditioning.

Bayes' rule provides the backward inference — the posterior distribution:

$$p(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$
$$= \frac{p(x, y)}{p(y)} = \frac{p(x)\, p(y \mid x)}{\sum_{x'} p(x')\, p(y \mid x')}.$$

The numerator $p(x)\, p(y \mid x)$ is the chain rule (prior $\times$ likelihood). The denominator $\sum_{x'} p(x')\, p(y \mid x')$ is the rule of total probability (a normalizing constant that ensures the posterior sums to 1).

1.8.5 Cause, Effect, and Conditioning

The two directions of conditional probability correspond to two fundamentally different operations:

(1) **Forward:** cause $\rightarrow$ effect, physical, given. fire $\rightarrow$ alarm. We observe the cause and predict the effect. This direction is often physically given — we know the mechanism.

(2) **Backward:** effect $\rightarrow$ cause, mental, inferred. alarm $\rightarrow$ fire. We observe the effect and infer the cause. This direction requires Bayes' rule.

The forward direction is the direction of nature: causes produce effects. The backward direction is the direction of reasoning: we observe effects and try to figure out the causes. Bayes' rule is the bridge between them.

1.8.6 Independence

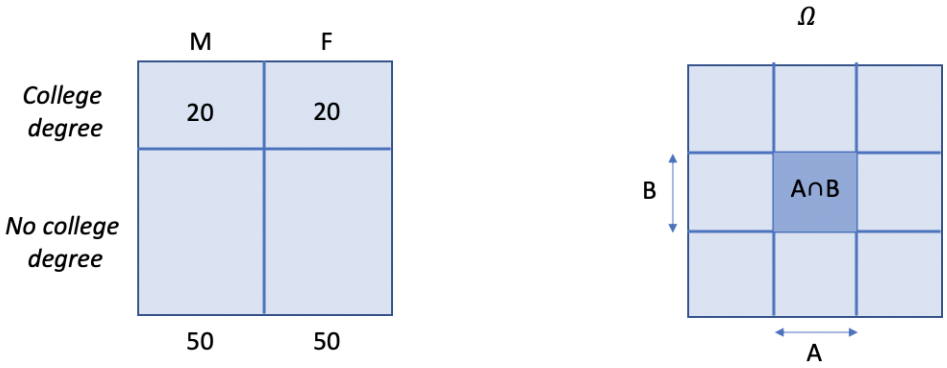


Figure 1.53: Independence: A and B have nothing to do with each other.

Two events A and B are **independent** if knowing that one has occurred gives no information about the other. There are two equivalent definitions:

Definition 1: $P(A \mid B) = P(A)$, equivalently $p(y \mid x) = p(y)$.

Definition 2: $P(A \cap B) = P(A)\,P(B)$, equivalently $p(x, y) = p(x)\,p(y)$.

Independence means: A and B (or X and Y) have nothing to do with each other. Learning that A occurred does not change the probability of B, and vice versa. In population language: the proportion of tall people is the same in the male sub-population as in the overall population.

Why are the two definitions equivalent? Starting from Definition 1: $P(A \mid B) = P(A)$. We know that $P(A \mid B) = P(A \cap B)/P(B)$. So $P(A) = P(A \cap B)/P(B)$, which gives $P(A \cap B) = P(A) \cdot P(B)$, which is Definition 2. The reasoning works in reverse as well.

1.8.7 Population of Sequences and Independence

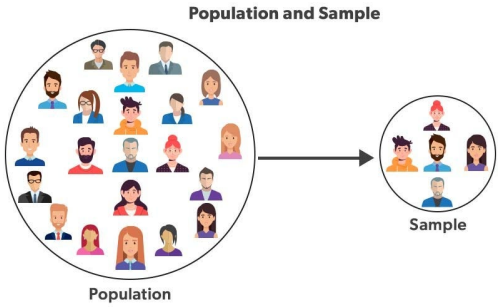


Figure 1.54: Population of sequences: equally likely outcomes + independent repetitions = equally likely sequences.

Sample a person from population Ω_1 of N people uniformly. Repeat n times independently. $\Omega_n = \{\text{all } N^n \text{ possible sequences}\}$.

Equally likely outcomes in Ω_1 + independent repetitions = equally likely sequences in Ω_n.

Let $\omega = (a_1, a_2, \ldots, a_n) \in \Omega_n$, each $a_i \in \Omega_1$. Then:

$$P(\omega) = P(a_1)\,P(a_2)\,\cdots\,P(a_n) = \frac{1}{N} \times \frac{1}{N} \times \cdots \times \frac{1}{N} = \frac{1}{N^n}.$$

This works for coin flipping ($\Omega_1 = \{\text{head, tail}\}$), die rolling ($\Omega_1 = \{1, 2, \ldots, 6\}$), or uniform random numbers ($\Omega_1 = [0, 1]$).

1.8.8 Conditional Independence

Independence is the simplest relationship between random variables: they have nothing to do with each other. But sometimes two random variables are *conditionally* independent — they are independent once we know the value of a third variable.

Markov chain: $C \to B \to A$, or $Z \to X \to Y$:

$$P(A \mid B, C) = P(A \mid B),$$
$$p(y \mid x, z) = p(y \mid x).$$

Future is independent of the past given the present. The immediate cause (parent) makes the remote cause (grandparent) irrelevant.

Meta rule: Insert the same condition in a definition or equation. For example, independence $P(A \cap B) = P(A)P(B)$ becomes conditional independence $P(A \cap C \mid B) = P(A \mid B)\,P(C \mid B)$.

Shared cause: $C \leftarrow B \to A$, or $X \leftarrow Z \to Y$:

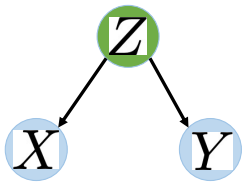


Figure 1.55: Shared cause: children are conditionally independent given parent.

$$P(A \cap C \mid B) = P(A \mid B)\,P(C \mid B),$$
$$p(x, y \mid z) = p(x \mid z)\,p(y \mid z).$$

Children are independent given the parent. For example, if Z = smoking, X = lung cancer, and Y = bronchitis, then lung cancer and bronchitis are not independent in the population (because smoking causes both, creating a correlation). But *given* whether a person smokes, lung cancer and bronchitis become independent — the shared cause has been accounted for.

1.8.9 Bayes Networks

A **Bayes network** (also called a directed acyclic graph or graphical model) is a compact representation of the joint distribution of many random variables, using the chain rule and conditional independence.

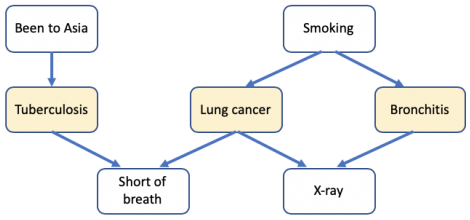


Figure 1.56: A Bayes network for medical diagnosis. a: been to Asia; s: smoking; t: tuberculosis; l: lung cancer; b: bronchitis; d: short of breath (dyspnea); x: X-ray.

The joint distribution factors according to the graph structure:

$$p(a, s, t, l, b, d, x) = p(a)\, p(s)\, p(t \mid a)\, p(l \mid s)\, p(b \mid s)\, p(d \mid t, l)\, p(x \mid b, l).$$

Each variable is conditionally independent of its non-descendants given its parents. This factorization dramatically reduces the number of parameters needed to specify the joint distribution.

Inference (e.g., computing $p(l \mid a, s, d, x)$) involves summing over hidden variables:

$$\begin{aligned} p(l \mid a, s, d, x) &= \frac{p(l, a, s, d, x)}{p(a, s, d, x)}, \\ p(l, a, s, d, x) &= \sum_{t,b} p(a, s, t, l, b, d, x), \\ p(a, s, d, x) &= \sum_{l} p(l, a, s, d, x). \end{aligned}$$

Efficient calculation: **message passing / belief propagation**. These algorithms exploit the graph structure to avoid summing over all possible combinations of hidden variables.

Two key conditional independence properties:

1. Sibling nodes are independent given their parent node.
2. A child node is independent of its grandparents given its parent.

1.8.10 Generative Pre-trained Transformer (GPT)

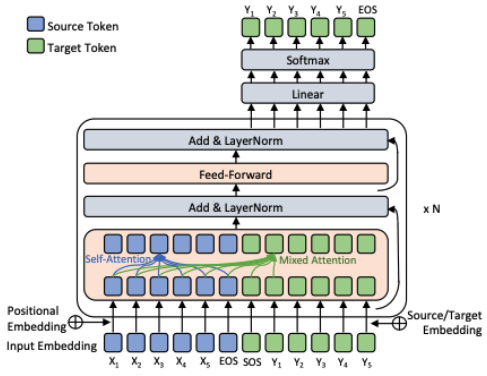


Figure 1.57: GPT: autoregressive generation of text.

A large language model like GPT generates text by modeling the conditional probability of each token given all previous tokens. This is simply the chain rule of probability applied to a sequence!

Let $x = (x_1, \ldots, x_{T_x})$ be a prompt (e.g., "Can you write a poem?") and $y = (y_1, \ldots, y_{T_y})$ be the response (e.g., "Certainly. Below is the poem..."):

$$p(y \mid x) = \prod_{t=1}^{T_y} p(y_t \mid y_{<t}, x).$$

This is simply the chain rule of probability applied to a sequence. Each word (or token) is generated one at a time, conditioned on all the previous words. The model learns the conditional probabilities from training data $(x^{(i)}, y^{(i)})_{i=1}^n$ by maximizing:

$$\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(y^{(i)} \mid x^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t} \log p_\theta(y_t^{(i)} \mid y_{<t}^{(i)}, x^{(i)}).$$

The model memorizes and generalizes (interpolation). The remarkable fact is that the mathematical foundation of the most advanced AI systems is nothing more than the chain rule and conditional probability that we have been studying in this chapter.

1.8.11 Denoising Diffusion Probability Model

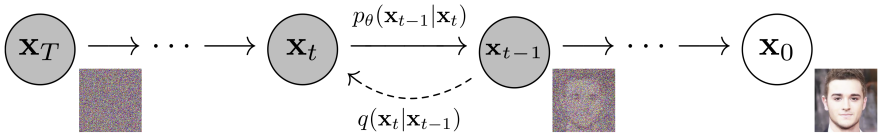


Figure 1.58: Diffusion model: forward noising and backward denoising.

Diffusion models generate images by reversing a noising process. x_0 is a clean image. We progressively add noise: $x_t = x_{t-1} + e_t$ with small noise e_t. The forward noising $q(x_t \mid x_{t-1})$ for $t = 1, \ldots, T$ turns the image into pure noise x_T. The backward denoising $p(x_{t-1} \mid x_t)$ is learned from data.

Think of it this way: the forward process is like gradually scrambling an image until it is unrecognizable white noise. The model learns to run this process backwards — starting from noise and gradually "denoising" it into a realistic image.

The model learns from training data $(x_0^{(i)})_{i=1}^n$ by maximizing:

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{t=T}^{1} \log p_\theta(x_{t-1}^{(i)} \mid x_t^{(i)}).$$

The model memorizes and generalizes (interpolation).

Both GPT and diffusion models are, at their core, applications of the chain rule and conditional probability. The deep mathematics of modern AI is built on the simple foundations we have developed in this chapter.

1.9 Take-Home Messages

Let us recap the main ideas of this chapter.

As long as you can count:
Count the population (of equally likely outcomes).
Count the repetitions (sequence of outcomes, fluctuation).
Population of sequences of repetitions (equally likely sequences).
Population of trajectories (random walk).
Two things:
(1) Intuition, visualization, and motivation.
(2) Precise notation and formula.
Accomplished: Most of the important concepts via intuitive examples.
Next: Systematic and more in-depth treatments — random variables and probability functions, expectation, continuous random variables, continuous-time processes.

Chapter 2

Random Variables

In Chapter 1, we introduced random variables informally as "numbers associated with outcomes." In this chapter, we study random variables systematically. We develop the machinery of probability distributions, expectation, and variance for both discrete and continuous random variables, and introduce the major probability distributions.

The key insight remains the same: a random variable X is a function defined on a population Ω of equally likely outcomes. The outcomes $\omega \in \Omega$ are equally likely, but the values $X(\omega)$ are generally *not* equally likely — some values are more common than others. The probability distribution of X describes how often each value occurs.

We shall study random variables more systematically, building the tools we need to analyze real-world phenomena.

2.1 Discrete Random Variables

Randomly sample a person ω from a population Ω of N people. Each person carries some attribute $X(\omega)$.

Connection to events:

$X(\omega)$: gender of ω, $\Omega \to \{0, 1\}$ — discrete. This random variable takes on only finitely many (or countably many) values.

$Y(\omega)$: height of ω, $\Omega \to \mathbb{R}^+$ — continuous. This random variable can take on any value in an interval.

$A = \{\omega : X(\omega) = 1\}$. $P(A) = P(X = 1)$.

$B = \{\omega : Y(\omega) > 6\}$. $P(B) = P(Y > 6)$.

$\omega \in \Omega$ equally likely, but $X(\omega)$ and $Y(\omega)$ are not necessarily equally likely. Think of it this way: every person in the population is equally likely to be sampled, but not every eye color is equally common. Some values of X might be shared by many people, making them more probable.

2.1.1 Probability Mass Function

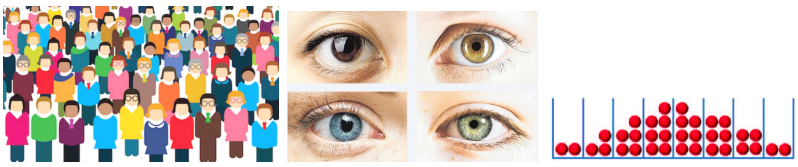


Figure 2.1: A population, eye color categories, and a histogram.

Example 2.1.1 (Eye color)**.** Let $X(\omega)$ be the eye color of person ω, coded as 1 (blue), 2 (brown), 3, 4. Let $N(x)$ be the number of people with eye color x.

The probability mass function describes the complete distribution of a discrete random variable. It tells us the probability of each possible value.

Definition 2.1.1 (Probability mass function)**.** The **probability mass function** (pmf), also called the **probability distribution** or **law**, of a discrete random variable X is:

$$X \sim p(x) = P(X = x) = \frac{N(x)}{N}.$$

The notation $X \sim p(x)$ means "X follows the distribution $p(x)$." The pmf $p(x)$ tells us, for each possible value x, what fraction of the population has that value. Since every person has exactly one eye color, the probabilities must sum to 1: $\sum_x p(x) = 1$.

We can organize the pmf in a table:

x	1	2	3	4
$p(x)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$
number	$N(1)$	$N(2)$	$N(3)$	$N(4)$

Example 2.1.2 (Number of siblings)**.** Let $X(\omega)$ be the number of siblings of person ω, taking values in $\{0, 1, 2, \ldots\}$.

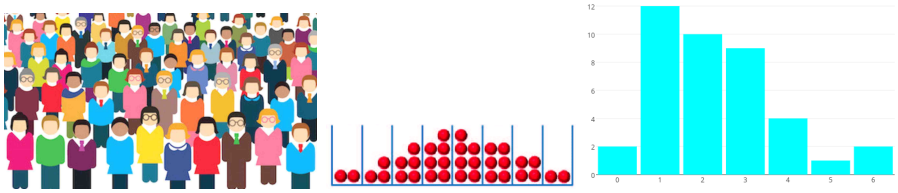


Figure 2.2: Population, histogram, and the sibling distribution.

$N(x)$ = number of people with x siblings. The pmf is:

$$X \sim p(x) = P(X = x) = \frac{N(x)}{N}, \qquad x = 0, 1, 2, \ldots$$

Unlike eye color, which has a finite number of categories, the number of siblings is a non-negative integer that can in principle be any value $0, 1, 2, \ldots$ (though very large values will have tiny probabilities). The histogram of the pmf gives a visual picture of how common each value is.

2.2 Expectation

2.2.1 Population Average

Now that we can describe the distribution of a random variable, we want to summarize it with a single number. The most natural summary is the **average**, which tells us the "center" or "typical value" of the distribution.

The **expectation** (or **expected value**, or **mean**) of a discrete random variable X is the population average of X. We derive it step by step, starting from the most basic definition and

simplifying:

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{N}\sum_{\omega\in\Omega} X(\omega) && \text{(average over all } N \text{ people)} \\
&= \frac{1}{N}\sum_{x} x\,N(x) && \text{(group by value: } N(x) \text{ people have value } x) \\
&= \sum_{x} x\,\frac{N(x)}{N} && \text{(factor out } 1/N) \\
&= \sum_{x} x\,p(x). && \text{(recognize } N(x)/N = p(x))
\end{aligned}$$

Let us make sure each step is clear. In the first line, we literally add up $X(\omega)$ for every person ω and divide by N — this is the plain average. In the second line, instead of going person by person, we group people by their value of X: there are $N(x)$ people with value x, and each contributes x to the sum, giving $x \cdot N(x)$ for that group. In the third line, we simply pull the $1/N$ inside the sum. In the fourth line, we recognize that $N(x)/N$ is exactly the probability $p(x)$.

The result is the "weighted average" formula: $\mathbb{E}(X) = \sum_x x\,p(x)$. Each value x is weighted by its probability $p(x)$. Values that are more probable contribute more to the average.

2.2.2 Long-Run Average: N^n Reasoning

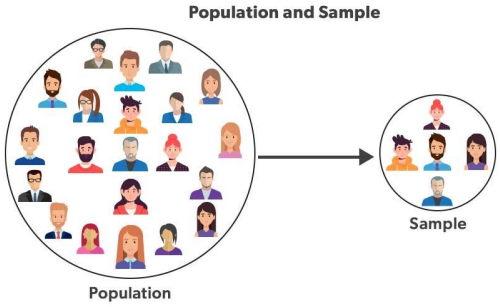


Figure 2.3: Repeat random sampling n times independently from a population.

The expectation also has a frequency interpretation, connecting it to what we observe in practice. Randomly sample a person ω from a population Ω of N people. Each person ω carries a number $X(\omega)$. The population average is $\mu = \mathbb{E}(X) = \sum_x x\,p(x)$.

Repeat random sampling n times independently $\to N^n$ equally likely sequences in the hyper-population Ω_n. For each sequence, compute $\bar{X}(\text{sequence}) = \text{average of the sequence}$.

The law of large numbers tells us:

$$\bar{X} \to \mu = \mathbb{E}(X) \quad \text{in probability as } n \to \infty.$$

More precisely, $A = \{\text{sequence} : |\bar{X}(\text{sequence}) - \mu| \le 0.01\}$ are the representative sequences, and $P(A) = |A|/N^n \to 1$ as $n \to \infty$.

In plain English: if you sample many people and average their values of X, the result will be close to $\mathbb{E}(X)$. The more people you sample, the closer the average gets. So the expectation $\mathbb{E}(X)$ is not just an abstract formula — it is the number that the sample average converges to.

2.2.3 Die Rolling: Bins on $[0, 1]$

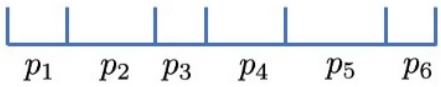


Figure 2.4: A non-uniform die: bins of different lengths on $[0, 1]$.

Here is a useful way to visualize a general discrete distribution. Imagine throwing a random point uniformly into $[0, 1]$. The interval is divided into six bins of possibly different lengths. The outcome $\omega \in \Omega = [0, 1]$ is equally likely (population of points, tiny balls), but $X(\omega)$ (the bin number) is not necessarily equally likely. The pmf is:

$$X \sim p(x) = P(X = x) = \text{length of bin } x.$$

The population average (each point or tiny ball ω carries a number $X(\omega)$) is $\mathbb{E}(X) = \sum_x x\, p(x)$.

Under independent repetitions, $p(x)$ is how often $X = x$ in the long run (e.g., throw 1 million points into $[0, 1]$).

x	1	2	3	4	5	6
$p(x)$	0.1	0.1	0.2	0.2	0.1	0.3

x	1	2	3	4	5	6
#	0.1m	0.1m	0.2m	0.2m	0.1m	0.3m
%	*10%*	*10%*	*20%*	*20%*	*10%*	*30%*

$$average = \frac{(1{\times}0.1m + 2{\times}0.1m + 3{\times}0.2m + 4{\times}0.2m + 5{\times}0.1m + 6{\times}0.3m)}{1m}$$

Figure 2.5: Long-run frequencies converge to probabilities.

2.2.4 Expectation of a Function

Sometimes we are interested not in X itself but in some function of X. For instance, if X is income, we might care about the tax $h(X)$, which is a nonlinear function of income. We need to compute the expected value of $h(X)$.

More generally, if h is any function, the expectation of $h(X)$ is:

$$\mathbb{E}(h(X)) = \sum_x h(x)\, p(x).$$

This is the population average (or long-run average) of $h(X)$: we apply h to each value x and weight by its probability. Notice that we do *not* need to first find the distribution of $h(X)$ and then compute its mean — we can compute $\mathbb{E}(h(X))$ directly from the distribution of X.

x	1	2	3	4	5	6	x
payoff	-$30	-$20	$0	$20	$30	$100	h(x)
	h(1)	*h(2)*	*h(3)*	*h(4)*	*h(5)*	*h(6)*	

$$longrun\ average = (-\$30)\times 0.1 + (-\$20)\times 0.1 + (\$0)\times 0.2 + (\$20)\times 0.2 + (\$30)\times 0.1 + (\$100)\times 0.3$$

Figure 2.6: Expectation of a function $h(X)$: each value $h(x)$ weighted by $p(x)$.

2.2.5 Utility

Offer 1	
x	$100
p(x)	1

$$E(X) = (\$100)\times 1 = \$100$$

Offer 2		
x	$0	$200
p(x)	1/2	1/2

$$E(X) = (\$0)\times\frac{1}{2} + (\$200)\times\frac{1}{2} = \$100$$

x: face value	$0	$100	$200
h(x): perceived value	$0	$100	$150

Figure 2.7: Two offers with different expected utilities.

The expectation of $h(X)$ has a natural interpretation as **expected utility**, reward, or value. Consider two offers:

Offer 1: $\mathbb{E}[h(X)] = \$100 \times 1 = \$100$.

Offer 2: $\mathbb{E}[h(X)] = \$0 \times 1/2 + \$150 \times 1/2 = \$75$.

If $h(x) = x$ (utility equals money), Offer 1 has higher expected utility. But if h is a nonlinear utility function — for example, if having \$100 is much more valuable to you than having \$0, but having \$150 is only slightly more valuable than \$100 — then the comparison might change. This connects to risk aversion, which we study in Chapter 4.

2.3 Variance

Expectation tells us the center of a distribution. But two distributions can have the same center and look very different — one might be tightly concentrated around the mean, while the other is spread out over a wide range. **Variance** tells us how spread out the distribution is — the extent of variation from the mean.

Definition 2.3.1 (Variance and standard deviation)**.** Let $\mu = \mathbb{E}(X)$. The **variance** of X is:

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \sum_x (x-\mu)^2\, p(x) = \sigma^2.$$

The **standard deviation** is $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sigma$.

The variance is the long-run average of the squared deviation from the mean. We square the deviations because otherwise positive and negative deviations would cancel out and the average deviation would be zero. The standard deviation has the same units as X (dollars, meters, etc.), while the variance has squared units (dollars^2, meters^2).

Example 2.3.1. Suppose X takes values \$0 and \$200, each with probability 1/2. Then $\mu = \mathbb{E}(X) = \$0 \times 1/2 + \$200 \times 1/2 = \$100$, and:

$$\mathrm{Var}(X) = (\$0 - \$100)^2 \times \frac{1}{2} + (\$200 - \$100)^2 \times \frac{1}{2} = \$^2\,10{,}000.$$

The standard deviation is $\mathrm{SD}(X) = \$100$.

More generally, $\mathrm{Var}(h(X)) = \mathbb{E}[(h(X) - \mathbb{E}[h(X)])^2]$.

2.3.1 Shortcut Formula

There is a convenient alternative formula for variance that often makes calculations easier. We derive it step by step:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[(X-\mu)^2] \\
&= \mathbb{E}[X^2 - 2\mu X + \mu^2] \qquad \text{(expand the square)} \\
&= \mathbb{E}(X^2) - 2\mu\,\mathbb{E}(X) + \mu^2 \qquad \text{(linearity of } \mathbb{E}) \\
&= \mathbb{E}(X^2) - 2\mu \cdot \mu + \mu^2 \qquad \text{(since } \mathbb{E}(X) = \mu) \\
&= \mathbb{E}(X^2) - \mu^2.
\end{aligned}$$

That is:

$$\boxed{\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.}$$

In words: the variance equals the "mean of the square minus the square of the mean." This formula is often easier to use because $\mathbb{E}(X^2) = \sum_x x^2\,p(x)$ is straightforward to compute — we just need the distribution of X, not the deviations from the mean.

The linearity of expectation used above is a crucial property. Let us verify it:

$$\begin{aligned}
\mathbb{E}[h(X) + g(X)] &= \sum_x [h(x) + g(x)]\,p(x) \\
&= \sum_x h(x)\,p(x) + \sum_x g(x)\,p(x) \\
&= \mathbb{E}[h(X)] + \mathbb{E}[g(X)].
\end{aligned}$$

We simply split the sum. The expectation of a sum is the sum of expectations. This "linearity of expectation" will be used over and over again throughout this book.

2.3.2 Connection to Data

If we sample $x_1, x_2, \ldots, x_n$ from $p(x)$ (e.g., rolling a die $\to$ 2, 1, 6, 5, 3, 2, 5, 4, 3, ...), then the sample statistics converge to the population parameters:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \to \mathbb{E}(X) = \mu, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \to \mathrm{Var}(X) = \sigma^2.$$

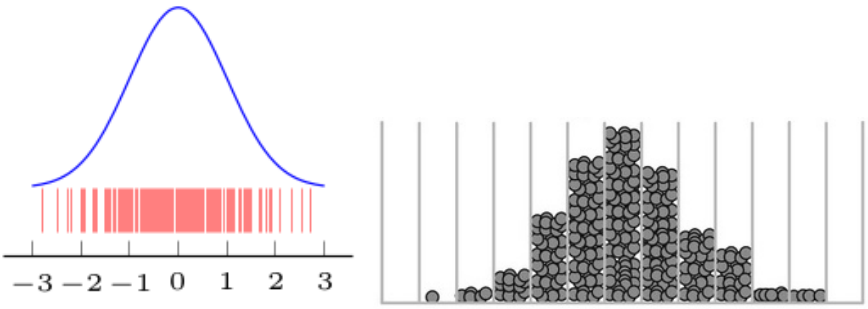


Figure 2.8: Sampling from a distribution: sample statistics converge to population parameters.

This is the law of large numbers in action. The sample mean $\bar{x}$ converges to the population mean μ, and the sample variance s^2 converges to the population variance σ^2. This is why we can learn about a population by studying a sample.

2.4 Linear Transformations

How do expectation and variance change when we apply a linear transformation $Y = aX + b$? This is a fundamental question because linear transformations are everywhere — converting between units (Celsius to Fahrenheit), standardizing data, and many other operations.

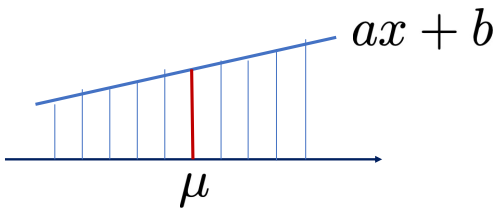


Figure 2.9: Linear transformation $Y = aX + b$.

2.4.1 Expectation under Linear Transformation

We derive the rule by direct computation:

$$\begin{aligned}
\mathbb{E}(Y) = \mathbb{E}(aX + b) &= \sum_x (ax + b)\, p(x) \\
&= \sum_x a\, x\, p(x) + \sum_x b\, p(x) \\
&= a \sum_x x\, p(x) + b \sum_x p(x) \\
&= a\, \mathbb{E}(X) + b.
\end{aligned}$$

The third step uses the fact that constants can be pulled out of sums. The last step uses $\sum_x p(x) = 1$ (total probability is 1). So the expectation of a linear transformation is the linear transformation of the expectation. This is not surprising — if you multiply everyone's salary by 2 and add 1000, the average salary gets multiplied by 2 and increased by 1000.

2.4.2 Variance under Linear Transformation

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathbb{E}[((aX+b)-\mathbb{E}(aX+b))^2] \\ &= \mathbb{E}[(aX+b-(a\mathbb{E}(X)+b))^2] \\ &= \mathbb{E}[(a(X-\mathbb{E}(X)))^2] \\ &= a^2\,\mathbb{E}[(X-\mathbb{E}(X))^2] = a^2\,\mathrm{Var}(X).\end{aligned}$$

Notice that the constant b disappears! Shifting a distribution (adding a constant) does not change its spread — it only moves the center. Scaling by a multiplies the variance by a^2. Taking the square root, the standard deviation gets multiplied by $|a|$.

So:

$$\boxed{\mathbb{E}(aX+b) = a\,\mathbb{E}(X)+b, \qquad \mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X).}$$

These formulas have direct counterparts in data. If $y_i = ax_i + b$, then $\bar{y} = a\bar{x} + b$ and:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i+b-(a\bar{x}+b))^2 = a^2\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

The data formula mirrors the population formula exactly.

2.5 Bernoulli and Binomial Distributions

We now study our first named distributions: the Bernoulli (a single coin flip) and the Binomial (many coin flips). These are the workhorses of discrete probability.

2.5.1 Bernoulli Distribution

The simplest non-trivial random variable is a coin flip.

$Z \sim \mathrm{Bernoulli}(p)$: $Z \in \{0,1\}$, $P(Z=1) = p$ and $P(Z=0) = 1-p$.

We compute the mean and variance. The mean is:

$$\mathbb{E}(Z) = 0\times(1-p) + 1\times p = p.$$

This makes intuitive sense: if the coin lands heads 30% of the time, the average number of heads per flip is 0.3.

For the variance, we compute it from the definition:

$$\begin{aligned}\mathrm{Var}(Z) &= (0-p)^2\times(1-p) + (1-p)^2\times p \\ &= p^2(1-p) + (1-p)^2 p = p(1-p)[p+(1-p)] = p(1-p).\end{aligned}$$

Alternatively, using the shortcut formula:

$$\mathbb{E}(Z^2) = 0^2\times(1-p) + 1^2\times p = p.$$

$$\mathrm{Var}(Z) = \mathbb{E}(Z^2) - [\mathbb{E}(Z)]^2 = p - p^2 = p(1-p).$$

Notice that $\mathrm{Var}(Z)$ is maximized at $p = 1/2$ (the most uncertain coin) and equals zero at $p = 0$ or $p = 1$ (a deterministic outcome has no variability).

2.5.2 Binomial Distribution

Flip a coin (p = probability of head) n times independently. X = number of heads.

$X \sim \text{Binomial}(n, p)$:

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

$\binom{n}{k}$ is the number of sequences with exactly k heads. $p^k(1-p)^{n-k}$ is the probability of each such sequence (since the flips are independent, we multiply the probabilities). For example, $n = 3$: $P(X = 2) = P(HHT) + P(HTH) + P(THH) = 3p^2(1-p)$.

When $p = 1/2$: $P(X = k) = \binom{n}{k}/2^n$, which is the fair coin case from Chapter 1.

2.5.3 Recall: Independence

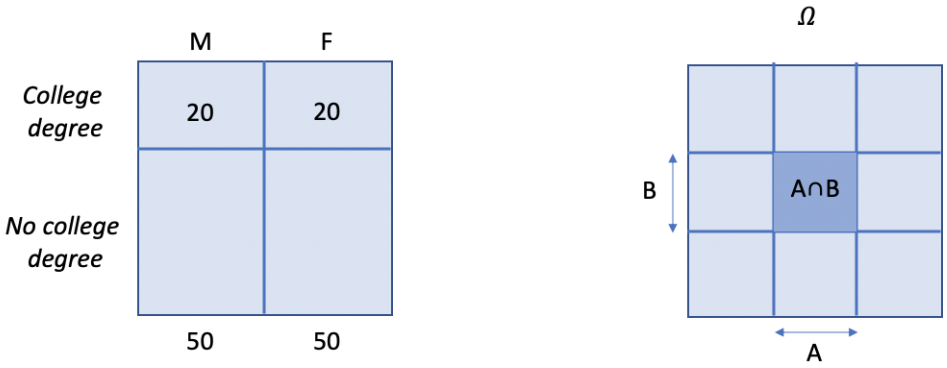


Figure 2.10: Independence: $P(A \cap B) = P(A)P(B)$.

The binomial formula relies on independence: $P(A \mid B) = P(A)$, equivalently $P(A \cap B) = P(A)P(B)$. Without independence, the probability of a specific sequence would not factor into a product of individual flip probabilities.

2.5.4 The Binomial Formula

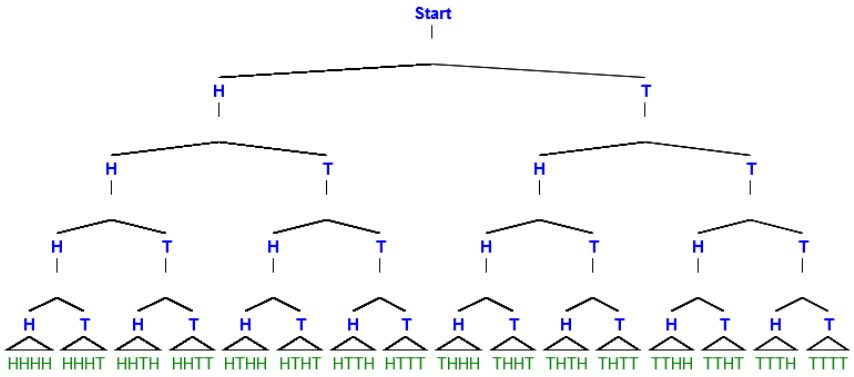


Figure 2.11: The tree of coin flip sequences.

The binomial formula can also be understood through algebraic expansion:

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

$n = 1$: $H + T$.

$n = 2$: $(H+T)(H+T) = HH + HT + TH + TT$.

$n = 3$: above $\times (H+T) = HHH + HHT + HTH + HTT + THH + THT + TTH + TTT$.

Each term in the expansion corresponds to one specific sequence of outcomes, and the binomial coefficient $\binom{n}{k}$ counts how many sequences have exactly k H's.

2.5.5 Binomial and Bernoulli: The Connection

The binomial random variable is simply the sum of independent Bernoulli random variables:

$X = Z_1 + Z_2 + \cdots + Z_n$, where $Z_i \sim \text{Bernoulli}(p)$ independently.

This connection allows us to compute the mean and variance easily:

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

Due to independence of Z_i, $i = 1, \ldots, n$ (which we will prove in Chapter 3 implies that Var of a sum equals the sum of Var's):

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(Z_i) = np(1-p).$$

Compare the simplicity of this approach to the direct calculation below!

2.5.6 Frequency and the Law of Large Numbers

X/n is the frequency of heads. Its mean and variance are:

$$\mathbb{E}(X/n) = \mathbb{E}(X)/n = p, \qquad \text{Var}(X/n) = \text{Var}(X)/n^2 = p(1-p)/n.$$

$\text{Var}(X/n) \to 0$ as $n \to \infty$, so $X/n \to p$ in probability. This is the **law of large numbers**. Probability = long-run frequency.

The key point: the variance of the frequency shrinks like $1/n$. This means the frequency becomes more and more concentrated around p as n grows. With 100 flips, the frequency might fluctuate between 0.4 and 0.6. With 10,000 flips, it will typically be between 0.49 and 0.51.

X/n is the average of Z_i. Probability p is the expectation of Z_i.

2.5.7 Direct Derivation of Binomial Expectation

We can also compute $\mathbb{E}(X)$ directly from the pmf, without using the Bernoulli decomposition. This derivation is more involved but illustrates a useful algebraic trick:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} k\, P(X=k) = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} np \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1}(1-p)^{n-k} \qquad \text{(pull out } np \text{ from } k\cdot n!/k!) \\
&= np \sum_{k'=0}^{n'} \binom{n'}{k'} p^{k'}(1-p)^{n'-k'} \qquad (k'=k-1,\ n'=n-1) \\
&= np \cdot 1 = np. \qquad \text{(the remaining sum is total probability = 1)}
\end{aligned}$$

The trick is in the second step: $k \cdot n!/(k!) = n \cdot (n-1)!/((k-1)!)$. This "peeling" trick — pulling one factor out of the factorial — is a standard technique for computing expectations of combinatorial distributions.

2.5.8 Direct Derivation of Binomial Variance

Similarly, we use the same peeling trick twice:

$$\begin{aligned}
\mathbb{E}(X(X-1)) &= \sum_{k=0}^{n} k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=2}^{n} n(n-1)p^2 \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2}(1-p)^{n-k} \qquad (k'=k-2,\ n'=n-2) \\
&= n(n-1)p^2 \sum_{k'=0}^{n'} \binom{n'}{k'} p^{k'}(1-p)^{n'-k'} = n(n-1)p^2.
\end{aligned}$$

Then we recover the variance:

$$\begin{aligned}
\mathbb{E}(X^2) &= \mathbb{E}(X(X-1)) + \mathbb{E}(X) = n(n-1)p^2 + np. \\
\mathrm{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = n(n-1)p^2 + np - (np)^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 = np - np^2 = np(1-p).
\end{aligned}$$

Both methods — the Bernoulli decomposition and the direct calculation — give $\mathrm{Var}(X) = np(1-p)$. The Bernoulli decomposition is much simpler, which demonstrates the power of decomposing a complex random variable into a sum of simpler ones.

2.6 Geometric Distribution

The geometric distribution models **waiting time**: how long do we have to wait until something happens for the first time?

$T \sim \text{Geometric}(p)$: T is the number of flips to get the first head, if we flip a coin independently with probability p of head in each flip.

$$P(T=k) = (1-p)^{k-1}p, \qquad k = 1, 2, 3, \ldots$$

For example: $T = 1$ (H) means the first flip is heads; $T = 2$ (TH) means the first flip is tails and the second is heads; $T = 3$ (TTH); $T = 4$ (TTTH). The probability of each outcome is the product of $(1 - p)$ for each tail followed by p for the final head. This models **waiting time**: how many trials until the first success?

2.6.1 Geometric Expectation

How long should we expect to wait? Let $q = 1 - p$. Then:

$$\mathbb{E}(T) = \sum_{k=1}^{\infty} k\, P(T = k) = \sum_{k=1}^{\infty} k\, q^{k-1} p = p \sum_{k=1}^{\infty} k\, q^{k-1}.$$

The key trick: recognize $k\, q^{k-1} = \frac{d}{dq} q^k$. This converts the sum into a derivative of a geometric series, which we know how to evaluate:

$$\mathbb{E}(T) = p \frac{d}{dq} \sum_{k=1}^{\infty} q^k = p \frac{d}{dq} \left(\frac{1}{1-q} - 1 \right) = p \cdot \frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

If $p = 1/6$ (rolling a specific number on a die), the expected waiting time is $1/p = 6$ rolls. This makes intuitive sense: if an event happens with probability $1/6$ on each trial, you should expect to wait about 6 trials.

2.6.2 Geometric Series

The computation above uses the geometric series. We derive it by a telescoping argument:

$$\begin{aligned}(1 - a)(1 + a + a^2 + \cdots + a^m) &= (1 + a + \cdots + a^m) - (a + a^2 + \cdots + a^{m+1}) \\ &= 1 - a^{m+1}.\end{aligned}$$

Most terms cancel (this is why we call it "telescoping"). Therefore: $1 + a + a^2 + \cdots + a^m = \frac{1 - a^{m+1}}{1 - a}$.

If $|a| < 1$, then $a^{m+1} \to 0$ as $m \to \infty$, giving $\sum_{k=0}^{\infty} a^k = \frac{1}{1-a}$. This infinite geometric series is one of the most useful formulas in mathematics.

2.6.3 Aside: The Quantum Bit

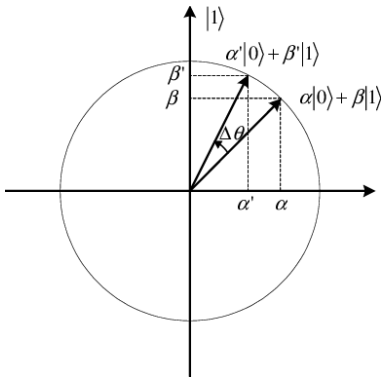


Figure 2.12: A quantum bit (qubit): a superposition of $|0\rangle$ and $|1\rangle$.

A classical bit is either 0 or 1. A quantum bit (qubit) has a state vector $\alpha|0\rangle + \beta|1\rangle$, which is a *superposition* of the two states. The state vector rotates over time (governed by Schrödinger's equation), and its squared length $|\alpha|^2 + |\beta|^2 = 1$ is preserved under rotation. When we observe (measure) the qubit, we get 0 with probability $p(0) = |\alpha|^2$ and 1 with probability $p(1) = |\beta|^2$.

Figure 2.13: Schrödinger's cat: $P(\text{alive}) = (1/\sqrt{2})^2 = 1/2$.

Before measurement, the qubit is in both states simultaneously. After measurement, it "collapses" to one state. The probabilities are determined by the amplitudes α and β, connecting quantum mechanics to our theory of probability.

2.7 Continuous Random Variables

So far, all our random variables have been discrete — they take on a finite or countable set of values. But many quantities in the real world are continuous: height, weight, temperature, time. In this section, we extend our framework to handle continuous random variables.

2.7.1 Density as a Limit

Randomly sample a person ω from a population Ω of N people. Let $X(\omega)$ be the height of person ω. We want to describe the density or distribution of these N points on the real line.

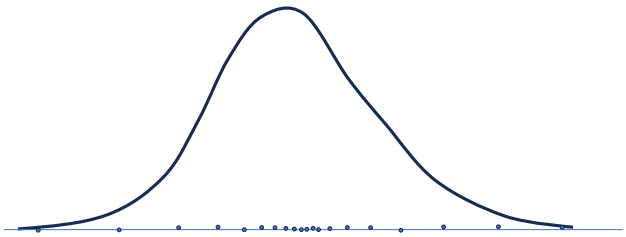


Figure 2.14: Population scatterplot: each person's height plotted as a point.

$N(x)$ = number of people in the bin $(x, x + \Delta x)$ (e.g., 6 ft to 6 ft 1 inch), where the precision is $\Delta x = 1$ inch. We cannot ask "how many people have height exactly 6.0000... feet?" because the answer would be zero. Instead, we ask how many people have heights in a small *interval*.

Definition 2.7.1 (Probability density function)**.** The **probability density function** (pdf) of a continuous random variable X is:

$$X \sim f(x) = \frac{P(X \in (x, x + \Delta x))}{\Delta x} = \frac{N(x)/N}{\Delta x}.$$

Mathematical idealization: $N \approx \infty$, $\Delta x \to 0$.

Notice the key difference from the pmf: the density $f(x)$ is *not* a probability. Rather, $f(x)\,\Delta x$ is the probability that X falls in the tiny interval $(x, x+\Delta x)$. The density $f(x)$ is probability per unit length. It can be greater than 1 (unlike a probability), but it is always non-negative.

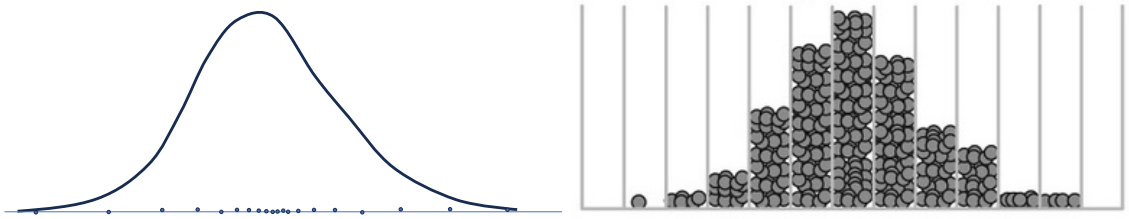


Figure 2.15: Left: population scatterplot. Right: histogram (bell curve).

Discretize the x-axis into equally spaced bins $(x, x+\Delta x)$. Then:

$$P(X \in (x, x+\Delta x)) = \frac{N(x)}{N} = f(x)\,\Delta x.$$

Here $f(x)$ is the height of the bin, and $f(x)\Delta x$ is the area. The total area under the density curve is 1:

$$\sum_x \frac{N(x)}{N} = \sum_x f(x)\,\Delta x \to \int f(x)\,dx = 1.$$

As $\Delta x \to 0$, the histogram becomes a smooth curve, and the sum becomes an integral.

2.7.2 Region Under the Curve

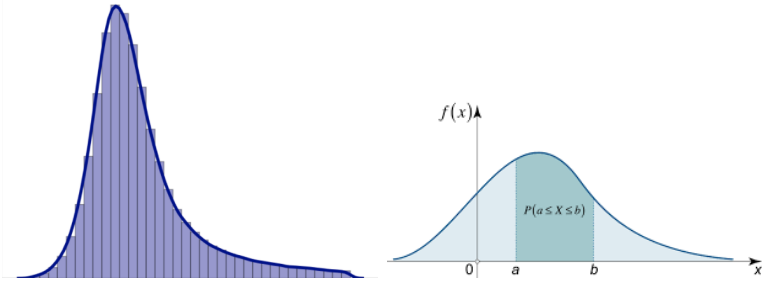


Figure 2.16: Left: region under the density curve (Ω). Right: probability as area.

There is a beautiful geometric picture for continuous probability. Randomly throw a point ω into the region Ω below curve $f(x)$. Ω is a population of points (tiny squares or balls). Let $X = X(\omega)$ be the horizontal coordinate of point ω. Then:

$$P(X \in (x, x+\Delta x)) = f(x)\,\Delta x.$$

The probability of X falling in an interval (a, b) is the area under the density curve between a and b:

$$P(X \in (a,b)) = \sum_{x \in (a,b)} f(x)\,\Delta x \to \int_a^b f(x)\,dx.$$

2.7.3 Independent Repetitions, Sample Scatterplot and Histogram

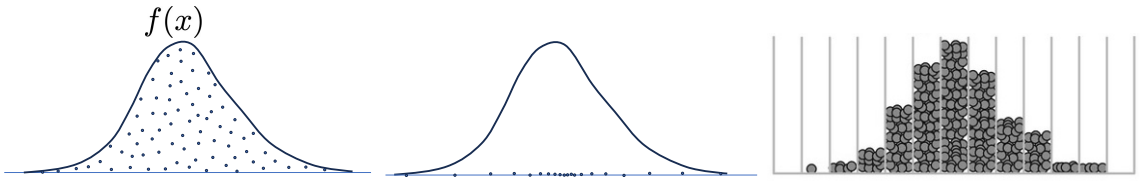


Figure 2.17: Left: sample scatterplot. Center: population scatterplot. Right: histogram.

Repeat n times, collapse to the x-axis, build a histogram. By N^n reasoning: sample scatterplot (random) $\approx$ population scatterplot (fixed), and sample histogram (random) $\approx$ population histogram (fixed). The sample histogram converges to the density curve as n grows — this is the continuous version of the law of large numbers.

2.7.4 Cumulative Distribution Function

Definition 2.7.2 (CDF)**.** The **cumulative distribution function** (CDF) is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

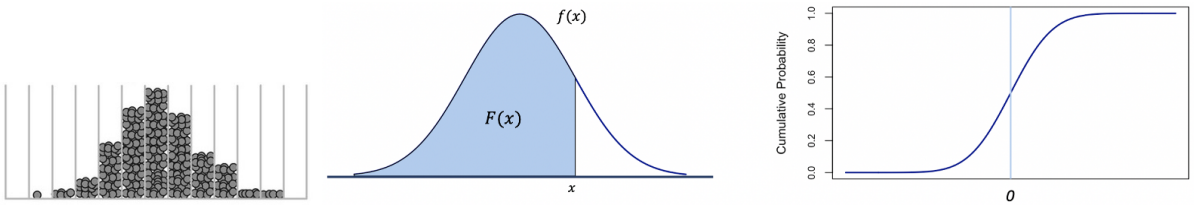


Figure 2.18: Density $f(x)$ and cumulative distribution function $F(x)$.

$F(x)$ tells us the percentage of people below x. For example, if X represents SAT scores and $F(1200) = 0.74$, it means 74% of test-takers scored 1200 or below — your score of 1200 puts you at the 74th percentile.

The density and CDF are related by:

Area: $F(x + \Delta x) - F(x) = f(x)\,\Delta x$. The probability of X falling in a tiny interval equals the density times the width.

Slope: $F'(x) = \frac{F(x+\Delta x)-F(x)}{\Delta x} = f(x)$. The density is the derivative of the CDF.

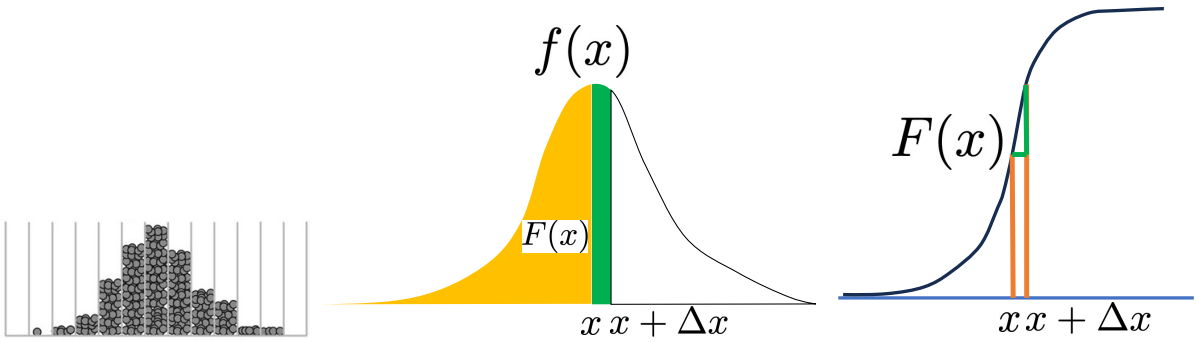


Figure 2.19: The density f is the slope of the CDF F. Notation: $dF(x) = F'(x)\,dx = f(x)\,dx$.

This relationship is just the fundamental theorem of calculus: the integral and derivative are inverse operations.

2.8 Expectation and Variance for Continuous Variables

2.8.1 Expectation

For a continuous random variable X with density $f(x)$, the expectation is defined in exact analogy with the discrete case. We simply replace sums with integrals.

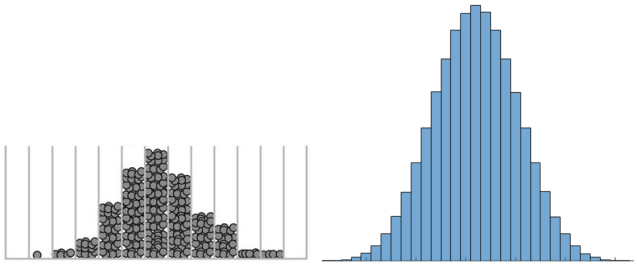


Figure 2.20: Density and histogram for expectation.

Recall discrete: $\mathbb{E}(X) = \sum x\,P(X = x) = \sum x\,p(x)$.
Continuous: $P(X \in (x, x + \Delta x)) = f(x)\,\Delta x$, so:

$$\mathbb{E}(X) = \sum x\,P(X \in (x, x + \Delta x)) = \sum x\,f(x)\,\Delta x \to \int x\,f(x)\,dx.$$

This is the population average, long-run average, and center.

As a population average, step by step:

$$\begin{aligned}\mathbb{E}(X) &= \frac{1}{N}\sum_{\omega} X(\omega) = \frac{1}{N}\sum_{x} x\,N(x) \\ &= \sum_{x} x\,\frac{N(x)}{N} = \sum_{x} x\,P(X \in (x, x + \Delta x)) \\ &= \sum_{x} x\,f(x)\,\Delta x \to \int x\,f(x)\,dx.\end{aligned}$$

More generally: $\mathbb{E}[h(X)] = \sum h(x)\,f(x)\,\Delta x \to \int h(x)\,f(x)\,dx$.

2.8.2 Variance

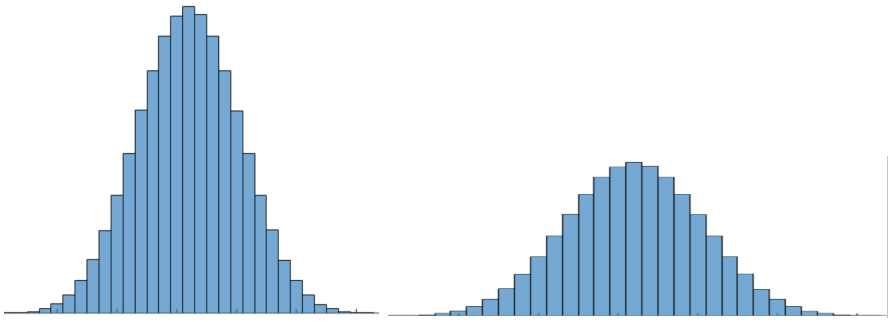


Figure 2.21: Density and variance: measuring fluctuation, volatility, spread.

$$\mathbb{E}(X) = \int x\, f(x)\, dx = \mu.$$

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2\, f(x)\, dx.$$

The shortcut formula still holds: $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$. The derivation is identical to the discrete case, using the linearity of the integral:

$$\begin{aligned}\mathbb{E}[r(X)+s(X)] &= \int [r(x)+s(x)]\, f(x)\, dx \\ &= \int r(x)\, f(x)\, dx + \int s(x)\, f(x)\, dx = \mathbb{E}[r(X)] + \mathbb{E}[s(X)].\end{aligned}$$

2.8.3 Linear Transformation (Continuous)

For $X \sim f(x)$, let $Y = aX + b$. The same formulas from the discrete case carry over:

$$\begin{aligned}\mathbb{E}(Y) = \mathbb{E}(aX+b) &= \int (ax+b)\, f(x)\, dx = a\int x\, f(x)\, dx + b\int f(x)\, dx = a\,\mathbb{E}(X) + b. \\ \mathrm{Var}(Y) = \mathrm{Var}(aX+b) &= \mathbb{E}[((aX+b) - \mathbb{E}(aX+b))^2] \\ &= \mathbb{E}[(aX+b-(a\mathbb{E}(X)+b))^2] \\ &= \mathbb{E}[a^2(X-\mathbb{E}(X))^2] = a^2\,\mathrm{Var}(X).\end{aligned}$$

2.8.4 Integration by Parts

Many expectation calculations for continuous distributions require **integration by parts**. Starting from the product rule for derivatives:

$$\frac{d}{dx}[u(x)v(x)] = u'(x)v(x) + u(x)v'(x),$$

equivalently $d(uv) = u\,dv + v\,du$.

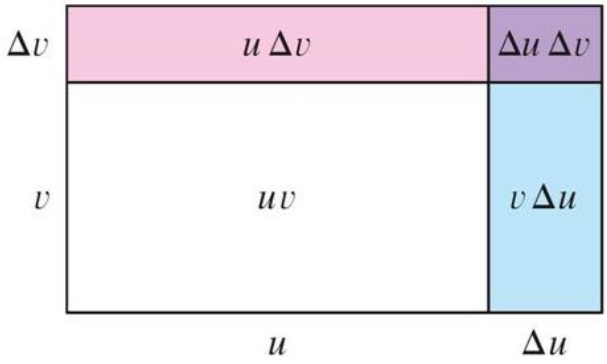


Figure 2.22: Geometric interpretation of integration by parts.

Integrate both sides:

$$\int [u'(x)v(x) + u(x)v'(x)]\,dx = u(x)v(x),$$

so:

$$\int u(x)\,v'(x)\,dx = u(x)v(x) - \int v(x)\,u'(x)\,dx, \qquad \text{i.e.,} \qquad \int u\,dv = uv - \int v\,du.$$

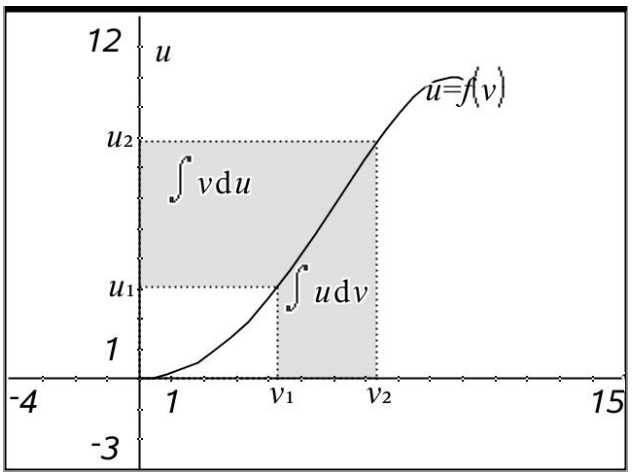


Figure 2.23: Integration by parts: trading $u\,dv$ for $v\,du$.

Notation: $\frac{du(x)}{dx} = \frac{d}{dx}u(x) = u'(x)$, so $du(x) = u'(x)\,dx$.

The idea of integration by parts is to trade one integral for a (hopefully simpler) one. We choose u and dv strategically so that $v\,du$ is easier to integrate than $u\,dv$.

2.9 The Exponential Distribution

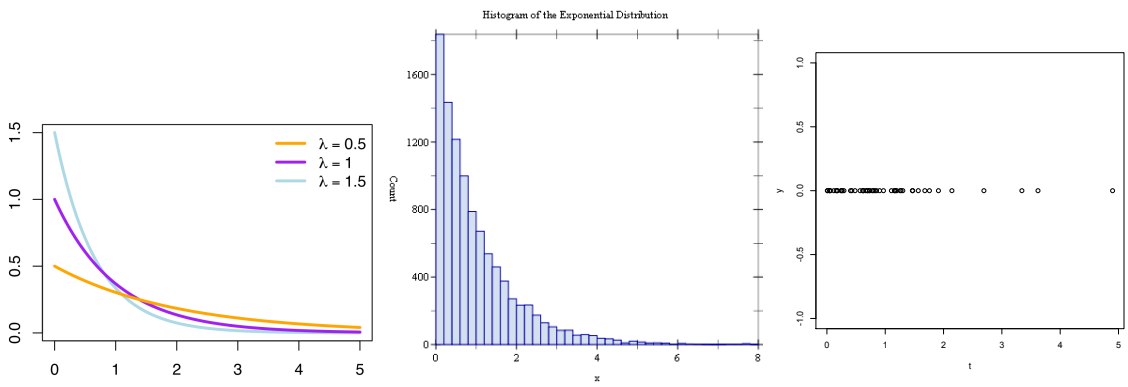


Figure 2.24: The exponential density, the decay process, and a random sample.

The exponential distribution is the continuous analogue of the geometric distribution — it models waiting time in continuous time. The parameter λ is the rate at which events occur.

$T \sim \text{Exponential}(\lambda)$: $f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$, $f(t) = 0$ for $t < 0$.

$P(T \in (t, t + \Delta t)) = \lambda e^{-\lambda t}\, \Delta t$.

Imagine 1 million particles; mark the times when they decay. These 1 million points on the real line have an exponential distribution. The number of points in $(t, t+\Delta t)$ is approximately $\lambda e^{-\lambda t}\, \Delta t$ million. Early on (small t), many particles decay. As time goes on, fewer and fewer remain, so the decay rate decreases.

2.9.1 CDF and Half-Life

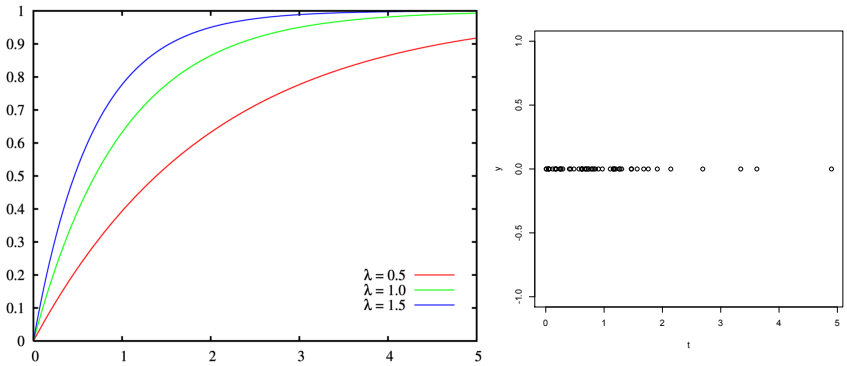


Figure 2.25: The exponential CDF: $F(t) = 1 - e^{-\lambda t}$.

$$F(t) = \int_0^t \lambda e^{-\lambda s}\, ds = -e^{-\lambda s}\Big|_0^t = 1 - e^{-\lambda t}.$$

$F(t)$ is the proportion of points below t — equivalently, the probability that a particle has decayed by time t.

The **half-life** t_{half} is the time by which half the particles have decayed:

$$F(t_{\text{half}}) = P(T \le t_{\text{half}}) = 1/2, \qquad t_{\text{half}} = \frac{\ln 2}{\lambda}.$$

2.9.2 Exponential Expectation

Using integration by parts with $u = t$ and $dv = \lambda e^{-\lambda t}\,dt$, so $du = dt$ and $v = -e^{-\lambda t}$:

$$\begin{aligned}
\mathbb{E}(T) &= \int_0^\infty t\,\lambda e^{-\lambda t}\,dt = -\int_0^\infty t\,d(e^{-\lambda t}) \\
&= -\left(t\,e^{-\lambda t}\Big|_0^\infty - \int_0^\infty e^{-\lambda t}\,dt \right) \\
&= -\left(0 - 0 + \frac{1}{\lambda} e^{-\lambda t}\Big|_0^\infty \right) \\
&= -\left(0 - \frac{1}{\lambda} \right) = \frac{1}{\lambda}.
\end{aligned}$$

If events occur at rate λ per unit time, the average wait is $1/\lambda$. This is the continuous analogue of the geometric expectation $\mathbb{E}(T) = 1/p$: rate and mean waiting time are reciprocals of each other.

2.10 The Normal (Gaussian) Distribution

The normal distribution is the most important distribution in probability and statistics. Its bell-shaped curve appears everywhere in nature: heights, test scores, measurement errors, stock returns, and much more. The reason for its ubiquity is the central limit theorem, which we will prove in Chapter 3.

2.10.1 Standard Normal

Let $Z \sim \mathcal{N}(0, 1)$, i.e., the density of Z is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

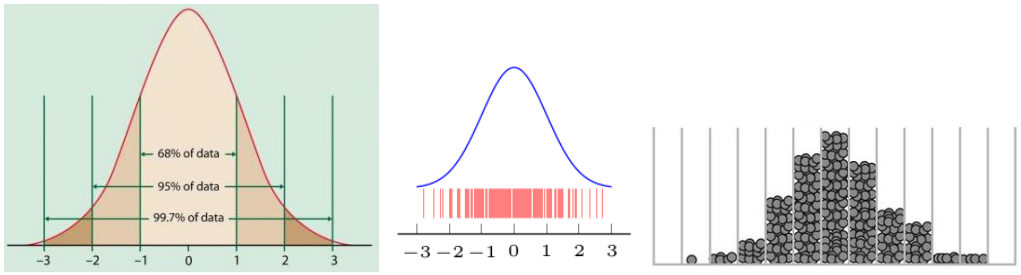


Figure 2.26: The standard normal density, data points, and the Galton board.

The bell-shaped curve is symmetric around 0. A key fact:

$$\int_{-2}^{2} f(z)\,dz = 95\%.$$

This means 95% of the area under the standard normal curve lies between -2 and 2. This is the famous "95% rule" and is one of the most frequently used facts in statistics.

2.10.2 Normal Expectation

$$\begin{aligned}\mathbb{E}(Z) &= \int_{-\infty}^{\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\ &= -\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\Big|_{-\infty}^{\infty} = 0.\end{aligned}$$

This follows because $d(e^{-z^2/2}) = -z\, e^{-z^2/2}\, dz$, so the integral is a total derivative that evaluates to zero at both limits. Alternatively, the integrand $z \cdot e^{-z^2/2}$ is an odd function (it changes sign when z is replaced by $-z$), so the integral over the symmetric interval $(-\infty, \infty)$ is zero. The density is symmetric around 0, so the mean must be 0.

2.10.3 Normal Variance

Using integration by parts: we write $z^2 e^{-z^2/2} = (-z) \cdot (-z\, e^{-z^2/2}) = (-z)\, d(e^{-z^2/2})/dz$, so:

$$\begin{aligned}\mathbb{E}(Z^2) &= \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-z)\, d(e^{-z^2/2}) \\ &= \frac{1}{\sqrt{2\pi}} \left(-z\, e^{-z^2/2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-z^2/2}\, d(-z) \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(0 - 0 + \int_{-\infty}^{\infty} e^{-z^2/2}\, dz \right) \\ &= \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = 1.\end{aligned}$$

Therefore $\mathrm{Var}(Z) = \mathbb{E}(Z^2) - [\mathbb{E}(Z)]^2 = 1 - 0 = 1$. The standard normal has mean 0 and variance 1 — that is precisely what "standard" means.

2.10.4 General Normal: Change of Variable

Let $Z \sim \mathcal{N}(0,1)$, and define $X = \mu + \sigma Z$, so $Z = (X - \mu)/\sigma$. This is a linear transformation, so we can use our formulas:

$$\mathbb{E}(X) = \mathbb{E}(\mu + \sigma Z) = \mu + \sigma\, \mathbb{E}(Z) = \mu.$$

$$\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2\, \mathrm{Var}(Z) = \sigma^2.$$

To find the density of X, we use the **change of density formula**. The idea is conservation of probability: the same number of people (or points) must be in corresponding intervals.

$$f(z)\, \Delta z = g(x)\, \Delta x.$$

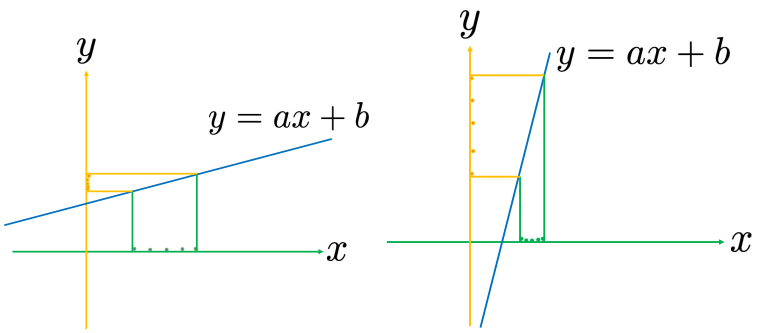


Figure 2.27: Change of density under linear transformation: space warping, stretching or squeezing.

For $y = ax + b$ (with $a > 0$), $x = (y - b)/a$, and $\Delta x/\Delta y = 1/a$, so:

$$g(y) = f(x)\,\frac{\Delta x}{\Delta y} = f\left(\frac{y-b}{a}\right)\cdot\frac{1}{a}.$$

Applying this to $X = \mu + \sigma Z$:

$$\begin{aligned} g(x) &= f(z)\,\frac{\Delta z}{\Delta x} = f\left(\frac{x-\mu}{\sigma}\right)\cdot\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \end{aligned}$$

$X \sim \mathcal{N}(\mu, \sigma^2)$ has density:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

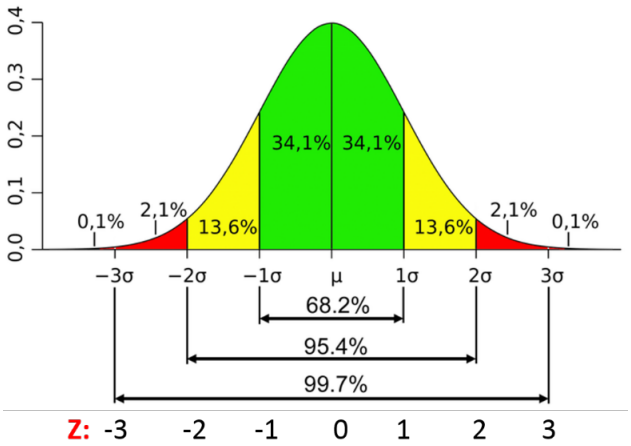


Figure 2.28: Normal densities with different means and variances.

Increasing μ shifts the bell curve to the right; increasing σ makes it wider and shorter (but the total area remains 1). The 95% rule generalizes: $P(\mu-2\sigma \le X \le \mu+2\sigma) = P(-2 \le Z \le 2) = 95\%$. About 95% of the data lies within 2 standard deviations of the mean.

Chapter 3

Two or More Random Variables

In the previous chapter, we studied a single random variable at a time. In practice, we almost always deal with multiple random variables simultaneously: a person's height *and* weight, a student's SAT score *and* GPA, a stock's price today *and* tomorrow. This chapter develops the machinery for handling two or more random variables: joint distributions, marginal and conditional distributions, covariance, correlation, and independence. We then use these tools to prove the two great limit theorems: the law of large numbers and the central limit theorem.

The key question throughout this chapter is: *how are two random variables related?* They might move together (positive correlation), move in opposite directions (negative correlation), or have no relationship at all (independence). Understanding these relationships is essential for prediction, inference, and decision-making.

3.1 Discrete Joint Distributions

3.1.1 Joint, Marginal, and Conditional

Consider a population of N people, each with two attributes: eye color X and hair color Y. We organize the population into a table, with rows indexed by X and columns indexed by Y.

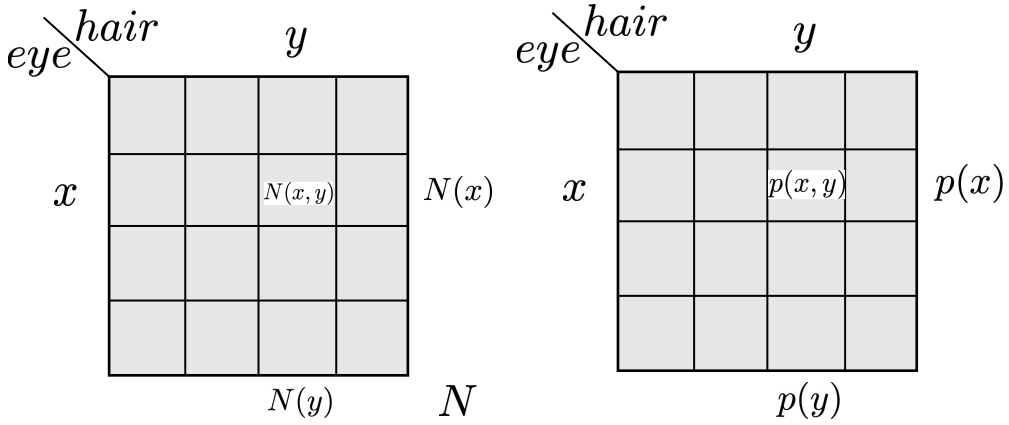


Figure 3.1: Left: counts $N(x, y)$. Right: probabilities $p(x, y) = N(x, y)/N$.

N: number of people in the population. $N(x, y)$: number of people with eye color x and hair color y. $N(x) = \sum_y N(x, y)$: number of people with eye color x (summing over all hair colors). $N(y) = \sum_x N(x, y)$: number of people with hair color y (summing over all eye colors).

The **joint distribution** describes the probability of each combination of values:

$$p(x, y) = \frac{N(x, y)}{N}.$$

The **marginal distributions** are obtained by summing over the other variable:

$$p(x) = \frac{N(x)}{N} = \frac{\sum_y N(x,y)}{N} = \sum_y p(x,y),$$
$$p(y) = \frac{N(y)}{N} = \frac{\sum_x N(x,y)}{N} = \sum_x p(x,y).$$

The name "marginal" comes from the fact that these are the row and column totals, written in the margins of the table. The marginal distribution of X tells us about X *ignoring* Y, and vice versa.

The **conditional distributions** are derived step by step:

$$p(x \mid y) = \frac{N(x,y)}{N(y)} = \frac{N(x,y)/N}{N(y)/N} = \frac{p(x,y)}{p(y)},$$
$$p(y \mid x) = \frac{N(x,y)}{N(x)} = \frac{N(x,y)/N}{N(x)/N} = \frac{p(x,y)}{p(x)}.$$

All of these are population proportions, and they approximate the corresponding sample proportions (frequencies) under repeated sampling: prob = population proportion $\approx$ sample proportion / frequency.

3.1.2 The Three Fundamental Rules

The three rules that govern all of probability are:

Marginalization: $p(y) = \sum_x p(x,y)$. To find the distribution of Y alone, sum over all possible values of X.

Conditioning (normalization): $p(x \mid y) = p(x,y)/p(y)$. To find the conditional distribution of X given $Y = y$, divide the joint by the marginal.

Chain rule (factorization): $p(x,y) = p(x)\,p(y \mid x)$. The joint distribution factors as the marginal of X times the conditional of Y given X.

These are not independent rules — each can be derived from the others. Together, they provide a complete calculus for manipulating joint distributions. Every computation in this chapter (and in much of probability and statistics) reduces to applying these three rules in the right order.

3.1.3 Cause and Effect: Forward and Backward

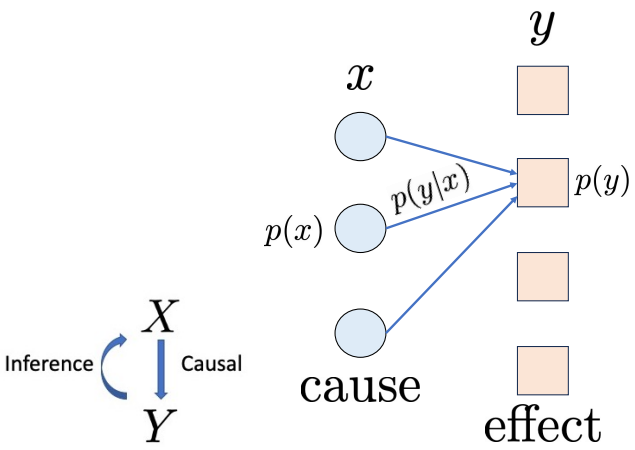


Figure 3.2: The cause-effect model: X causes Y.

When we think of X as a cause and Y as an effect, the three rules take on special meaning:

Marginal (prior): $p(x) = P(X = x)$, marginal $p(y) = P(Y = y)$. These describe each variable in isolation.

Forward conditional (generation): $p(y \mid x) = P(Y = y \mid X = x)$ — from cause to effect. This tells us how the cause produces the effect. It is the "mechanism."

Backward conditional (inference): $p(x \mid y) = P(X = x \mid Y = y)$ — from effect to cause. This tells us what we can infer about the cause from observing the effect.

Chain rule: Joint $p(x, y) = p(x)\, p(y \mid x)$. The joint probability of cause and effect is the prior probability of the cause times the forward conditional.

Rule of total probability: Marginal $p(y) = \sum_x p(x, y) = \sum_x p(x)\, p(y \mid x)$. The overall probability of the effect is obtained by averaging the forward conditional over all possible causes.

3.1.4 Connection to GPT

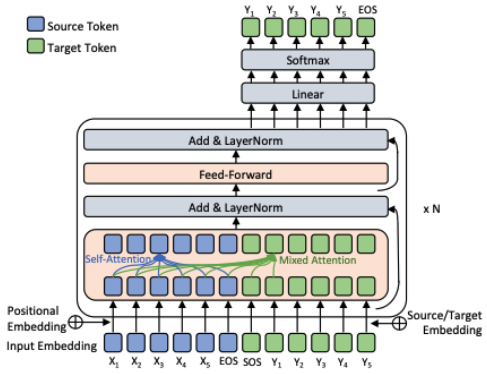


Figure 3.3: GPT: a chain rule decomposition of conditional text generation.

The chain rule is the mathematical foundation of large language models. Given a prompt $x = (x_1, \ldots, x_{T_x})$ (e.g., "Can you write a poem?"), GPT generates a response $y = (y_1, \ldots, y_{T_y})$ (e.g., "Certainly. Below is the poem..") via:

$$p(y \mid x) = \prod_{t=1}^{T_y} p(y_t \mid y_{<t}, x).$$

Each factor is a conditional probability, and the product is the chain rule applied to a sequence. The model learns these conditionals from data by maximizing:

$$\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(y^{(i)} \mid x^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t} \log p_\theta(y_t^{(i)} \mid y_{<t}^{(i)}, x^{(i)}).$$

The model memorizes and generalizes (interpolation).

3.1.5 Bayes' Rule Revisited

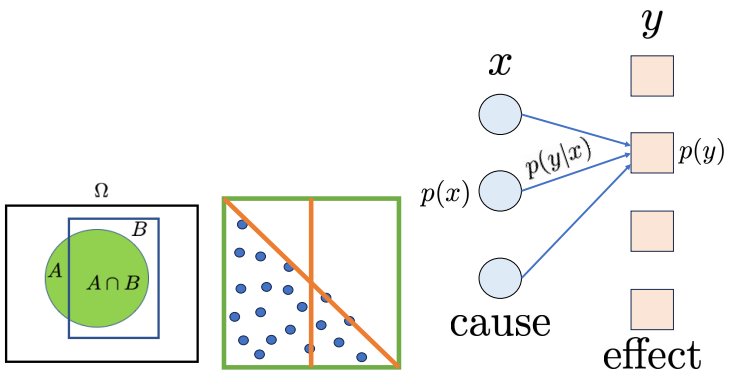


Figure 3.4: Bayes' rule: from forward to backward conditioning.

$P(A \mid B) = P(A \cap B)/P(B)$.

Bayes' rule provides backward inference, back tracing, and the posterior distribution:

$$p(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$
$$= \frac{p(x, y)}{p(y)} = \frac{p(x)\, p(y \mid x)}{\sum_{x'} p(x')\, p(y \mid x')}.$$

The numerator is the chain rule (prior $\times$ likelihood). The denominator is the rule of total probability. Bayes' rule is simply the conditioning rule applied twice — once to define the backward conditional, and once to compute it via the forward conditional.

3.1.6 Expectation for Joint Distributions

If (X, Y) has joint distribution $p(x, y)$, then the expectation of any function $h(X, Y)$ is:

$$\mathbb{E}[h(X, Y)] = \sum_{x,y} h(x, y)\, p(x, y).$$

This is the population average or long-run average. In terms of counting:

$$\frac{1}{N} \sum_{x,y} h(x, y)\, N(x, y) = \sum_{x,y} h(x, y)\, \frac{N(x, y)}{N} = \sum_{x,y} h(x, y)\, p(x, y) = \mathbb{E}[h(X, Y)].$$

As a special case, the expectation of X depends only on the marginal of X:

$$\mathbb{E}(X) = \sum_{x,y} x\, p(x, y) = \sum_{x} x \sum_{y} p(x, y) = \sum_{x} x\, p(x).$$

This is reassuring: to compute $\mathbb{E}(X)$, we only need the distribution of X, not the joint distribution of (X, Y).

Same for $\mathbb{E}[h(X)]$. The variance is defined as usual: $\mathrm{Var}(h(X, Y)) = \mathbb{E}[(h(X, Y) - \mathbb{E}[h(X, Y)])^2]$.

3.2 Continuous Joint Distributions

3.2.1 Joint Density

For two continuous random variables, X = height and Y = weight, the joint density is:

$$f(x,y) = \frac{P(X \in (x, x+\Delta x),\, Y \in (y, y+\Delta y))}{\Delta x\, \Delta y} = \frac{N(x,y)/N}{\Delta x\, \Delta y}.$$

The general principle is: **density = probability / size**. In one dimension, the size is Δx; in two dimensions, $\Delta x\, \Delta y$. The density tells us how many people (per unit area) are near the point (x, y).

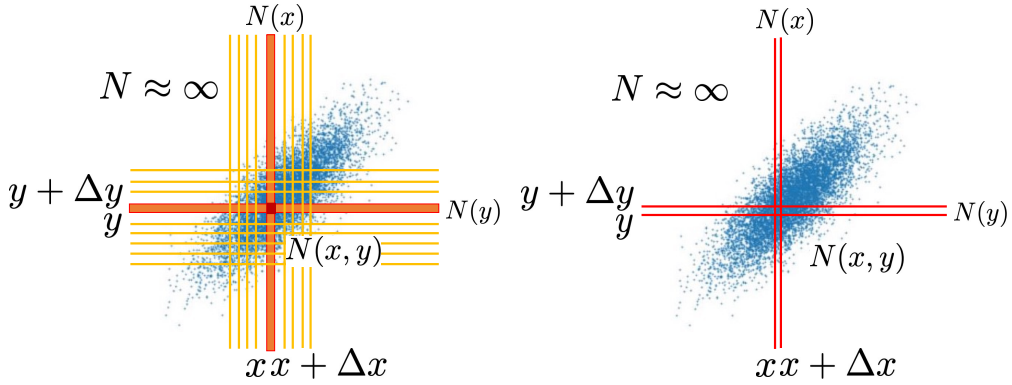


Figure 3.5: Left: discretized grid for two continuous variables. Right: joint density surface.

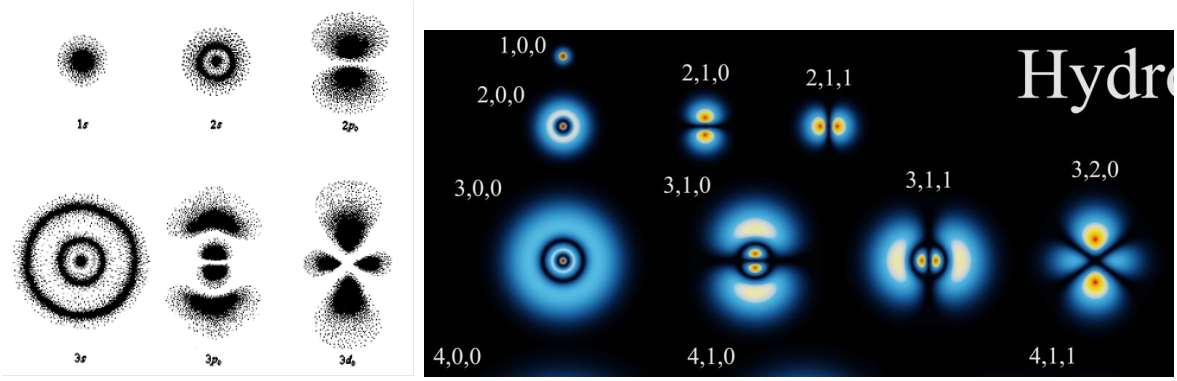


Figure 3.6: Probability density function as a heat map (electron cloud around nucleus). Population of N equally likely possibilities. Mathematical idealization: $N \approx \infty$. Density = prob mass in the cell / volume of cell.

3.2.2 Marginal Density

The marginal density of X is obtained by integrating out Y. The derivation uses the same logic as the discrete case, step by step:

$$f(x) = \frac{P(X \in (x, x+\Delta x))}{\Delta x} = \frac{N(x)/N}{\Delta x}$$
$$= \frac{\sum_y N(x,y)/N}{\Delta x} = \frac{\sum_y f(x,y)\, \Delta x\, \Delta y}{\Delta x} = \int f(x,y)\, dy.$$

Similarly, $f(y) = \int f(x, y)\, dx$.

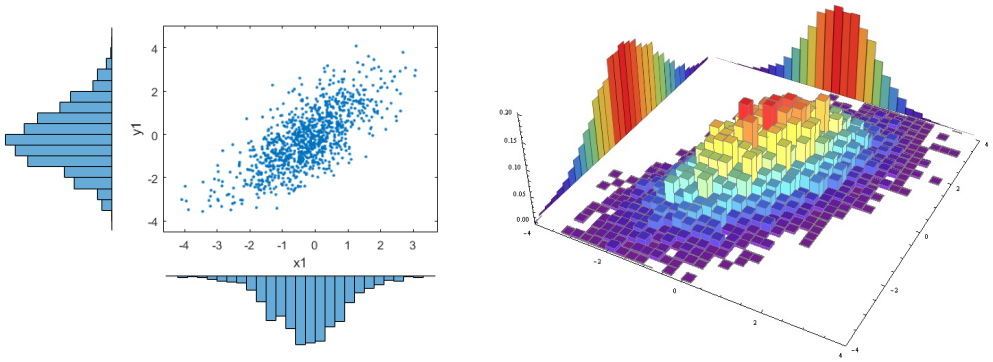


Figure 3.7: Joint density surface and its marginal projections.

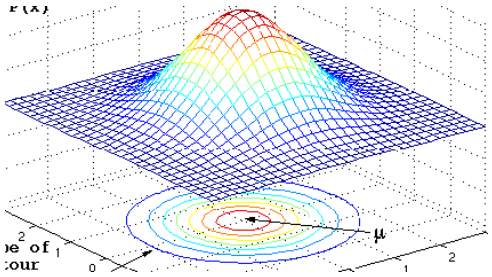


Figure 3.8: Sample points under the density surface, collapsed onto the plane.

Geometrically, integrating out Y is like "collapsing" the 2D density surface onto the X-axis. The resulting marginal density $f(x)$ describes the distribution of X alone, ignoring Y.

3.2.3 Conditional Density

The conditional density of Y given $X = x$ is derived by the same ratio logic:

$$\begin{aligned} f(y \mid x) &= \frac{P(Y \in (y, y + \Delta y) \mid X \in (x, x + \Delta x))}{\Delta y} \\ &= \frac{N(x, y)/N(x)}{\Delta y} = \frac{N(x, y)/N}{(N(x)/N)\, \Delta y} \\ &= \frac{f(x, y)\, \Delta x\, \Delta y}{f(x)\, \Delta x\, \Delta y} = \frac{f(x, y)}{f(x)}. \end{aligned}$$

Similarly, $f(x \mid y) = f(x, y)/f(y)$.

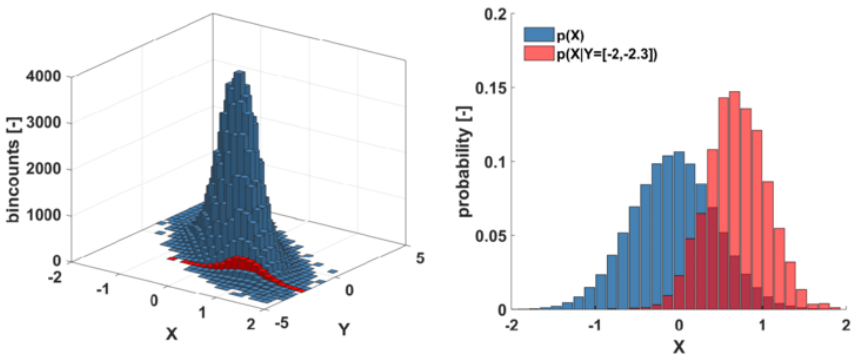


Figure 3.9: Conditional density: the shape of the density surface at a fixed value of x.

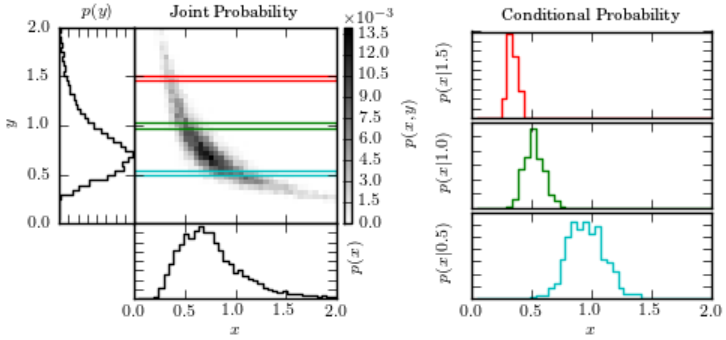


Figure 3.10: Conditional density slices at different values of x.

Think of slicing the joint density surface with a vertical plane at $X = x$. The shape of the cross-section (after normalization so the area is 1) is the conditional density $f(y \mid x)$. Different values of x give different conditional densities, reflecting how the distribution of Y changes as we condition on different values of X.

3.2.4 Rules for Continuous Distributions

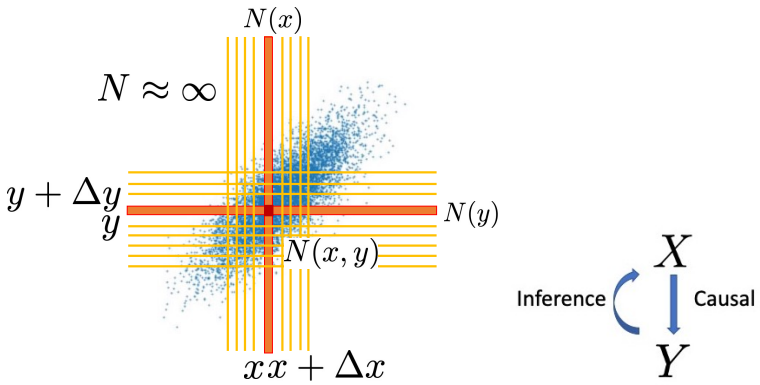


Figure 3.11: The same three rules apply in the continuous case.

Marginalization: $f(y) = \int f(x, y)\, dx$.

Normalization (conditioning): $f(x \mid y) = f(x, y)/f(y)$.

Factorization (chain rule): $f(x, y) = f(x)\, f(y \mid x)$.

$f(y \mid x)$: prediction. $f(x \mid y)$: inference. The rules are identical to the discrete case, with sums replaced by integrals.

3.2.5 Connection to Diffusion Models

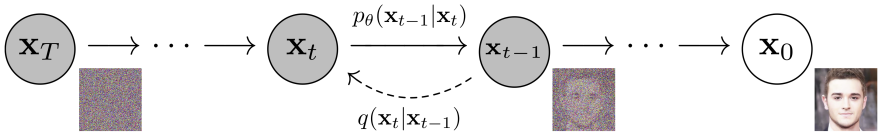


Figure 3.12: Diffusion model: the chain rule governs both forward noising and backward denoising.

x_0: clean image. $x_t = x_{t-1} + e_t$, with small noise e_t. Forward noising $q(x_t \mid x_{t-1})$ for $t = 1, \ldots, T$ turns image into big noise x_T. Backward denoising $p(x_{t-1} \mid x_t)$ is learned from training data $(x_0^{(i)})_{i=1}^n$ by maximizing:

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{t=T}^{1} \log p_\theta(x_{t-1}^{(i)} \mid x_t^{(i)}).$$

The model memorizes and generalizes (interpolation). The chain rule of probability governs both the forward noising process and the backward denoising process.

3.2.6 Expectation for Continuous Joint Distributions

If $(X, Y) \sim f(x, y)$:

$$\mathbb{E}[h(X, Y)] = \iint h(x, y)\, f(x, y)\, dx\, dy.$$

This is the population average or long-run average of $h(X, Y)$:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} h(X_i, Y_i) &= \frac{1}{n} \sum_{\text{cells}} h(x, y)\, n\, f(x, y)\, \Delta x\, \Delta y \\ &\to \iint h(x, y)\, f(x, y)\, dx\, dy. \end{aligned}$$

3.2.7 Conditional Expectation and Regression

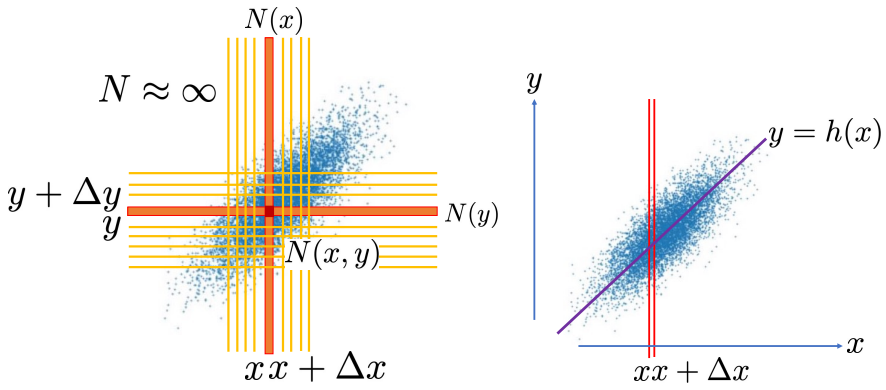


Figure 3.13: Conditional expectation $\mathbb{E}[Y \mid X = x]$: the mean of Y within a vertical slice.

Recall $\mathbb{E}(Y) = \int y\, f(y)\, dy$. The **conditional expectation** of Y given $X = x$ is:

$$h(x) = \mathbb{E}[Y \mid X = x] = \int y\, f(y \mid x)\, dy.$$

This is a **regression function** or **prediction**: the average value of Y among all individuals with $X = x$. If X is the number of hours a student studies, and Y is their exam score, then $\mathbb{E}[Y \mid X = x]$ predicts the expected score for a student who studies x hours.

The conditional variance measures the remaining uncertainty — how much Y varies even among people with the same value of X:

$$\mathrm{Var}(Y \mid X = x) = \mathbb{E}[(Y - h(x))^2 \mid X = x] = \int (y - h(x))^2\, f(y \mid x)\, dy.$$

3.3 The Bivariate Normal Distribution

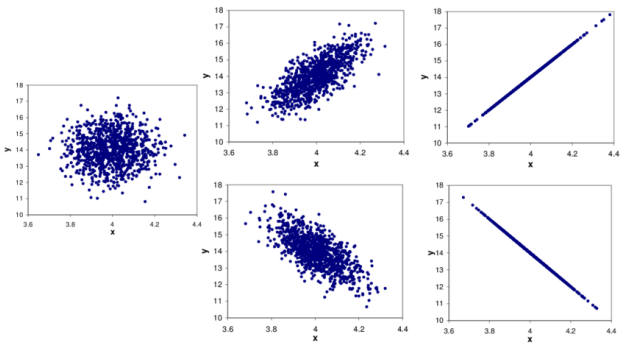


Figure 3.14: The bivariate normal: a bell-shaped surface in two dimensions.

The bivariate normal is the most important two-dimensional distribution. It has an elegant construction that makes its properties easy to derive:

$$X \sim \mathcal{N}(0,1),$$
$$Y = \rho X + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1-\rho^2), \quad |\rho| \le 1,$$

where ϵ is independent of X. The parameter ρ controls the correlation between X and Y. When $\rho = 0$, $Y = \epsilon$ is independent of X. When $\rho = 1$, $Y = X$ perfectly. When $\rho = -1$, $Y = -X$ perfectly. For values in between, there is a linear relationship with some noise.

3.3.1 Conditional Distribution

Given $X = x$, we have $Y = \rho x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1-\rho^2)$ is independent of X. Therefore:

$$\mathbb{E}(Y \mid X = x) = \mathbb{E}(\rho x + \epsilon) = \rho x.$$

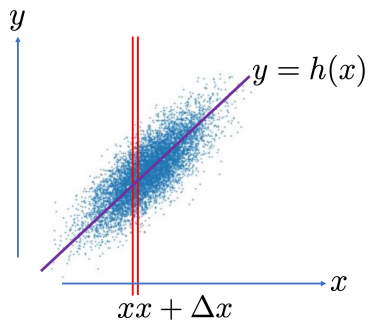


Figure 3.15: The distribution of points within a vertical slice at x.

This is **regression toward the mean**: since $|\rho| < 1$, the prediction ρx is closer to zero than x itself. For example, if X is a father's height (standardized) and Y is the son's height, then sons of unusually tall fathers tend to be tall but not *as* extreme as their fathers. This phenomenon was first observed by Francis Galton in the 1880s, and it is where the term "regression" comes from.

The conditional variance is:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{Var}(\rho x + \epsilon) = \mathrm{Var}(\epsilon) = 1 - \rho^2.$$

So $[Y \mid X = x] \sim \mathcal{N}(\rho x, 1-\rho^2)$. Notice that the conditional variance does not depend on x — the spread of Y around its conditional mean is the same for all values of X.

3.3.2 Joint Density

Using the chain rule $f(x,y) = f(x)\, f(y \mid x)$, step by step:

$$f(x,y) = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}_{f(x)} \cdot \underbrace{\frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-(y-\rho x)^2/[2(1-\rho^2)]}}_{f(y|x)}$$
$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right].$$

This density is symmetric in (x,y), which means $[X \mid Y = y] \sim \mathcal{N}(\rho y, 1-\rho^2)$ as well. The symmetry is remarkable: even though we constructed Y from X, the backward conditional has the same form as the forward conditional.

3.4 Covariance

We need a number that measures how two random variables move together. That number is the covariance.

Definition 3.4.1 (Covariance)**.** Let $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$. The **covariance** of X and Y is:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

It is defined for both discrete and continuous random variables.

The covariance measures how X and Y *co-vary* — whether they tend to be above their means at the same time (positive covariance) or whether one tends to be above its mean when the other is below (negative covariance).

Given data (X_i, Y_i), $i = 1, \ldots, n$, the sample covariance is:

$$\text{Cov}(X, Y) \approx \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

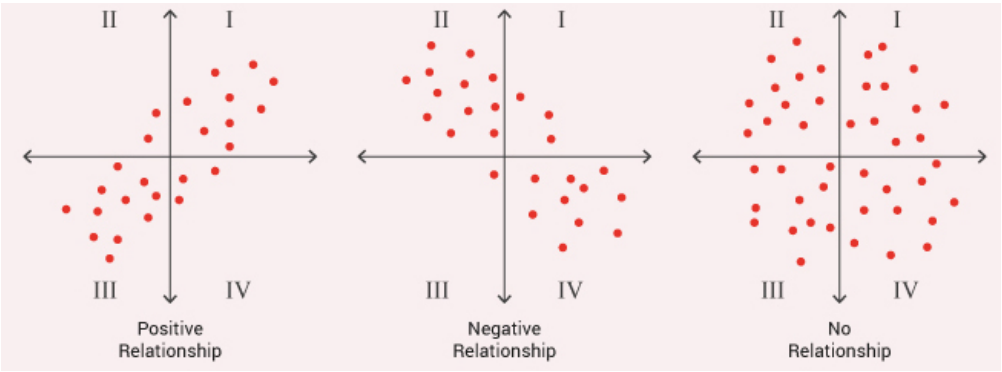


Figure 3.16: Covariance via quadrants. In quadrants I and III, $(X_i - \bar{X})(Y_i - \bar{Y}) > 0$. In quadrants II and IV, $(X_i - \bar{X})(Y_i - \bar{Y}) < 0$.

Divide the scatter plot into four quadrants centered at $(\bar{X}, \bar{Y})$. Points in quadrants I and III (both above mean, or both below mean) contribute positive products. Points in quadrants II and IV (one above, one below) contribute negative products. If the positive contributions outweigh the negative ones, the covariance is positive, indicating that X and Y tend to move together.

3.4.1 Shortcut Formula

The shortcut formula is derived step by step, exactly as for variance:

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathbb{E}[XY - \mu_X Y - X\mu_Y + \mu_X \mu_Y] \qquad \text{(expand)} \\
&= \mathbb{E}(XY) - \mu_X \, \mathbb{E}(Y) - \mu_Y \, \mathbb{E}(X) + \mu_X \mu_Y \qquad \text{(linearity of } \mathbb{E}) \\
&= \mathbb{E}(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \qquad \text{(substitute)} \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}$$

Clearly, $\text{Cov}(X, X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \text{Var}(X)$. The variance is the covariance of X with itself.

3.4.2 Linearity of Covariance

First property — scaling:

$$\begin{aligned}\mathrm{Cov}(aX+b,\, cY+d) &= \mathbb{E}[(aX+b-\mathbb{E}(aX+b))(cY+d-\mathbb{E}(cY+d))] \\ &= \mathbb{E}[a(X-\mathbb{E}(X))\, c(Y-\mathbb{E}(Y))] = ac\,\mathrm{Cov}(X,Y).\end{aligned}$$

Covariance depends on units (meters vs. feet, kilograms vs. pounds). If you measure height in centimeters instead of meters, the covariance with weight changes by a factor of 100.

Second property — additivity:

$$\begin{aligned}\mathrm{Cov}(X+Y,\, Z) &= \mathbb{E}[(X+Y-\mathbb{E}(X+Y))(Z-\mathbb{E}(Z))] \\ &= \mathbb{E}[(X-\mathbb{E}(X)+Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))] \\ &= \mathbb{E}[(X-\mathbb{E}(X))(Z-\mathbb{E}(Z))] + \mathbb{E}[(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))] \\ &= \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z).\end{aligned}$$

This additivity property will be essential when we compute the variance of sums.

3.5 Correlation

To remove the dependence on units, we standardize. The result is a dimensionless measure of linear association.

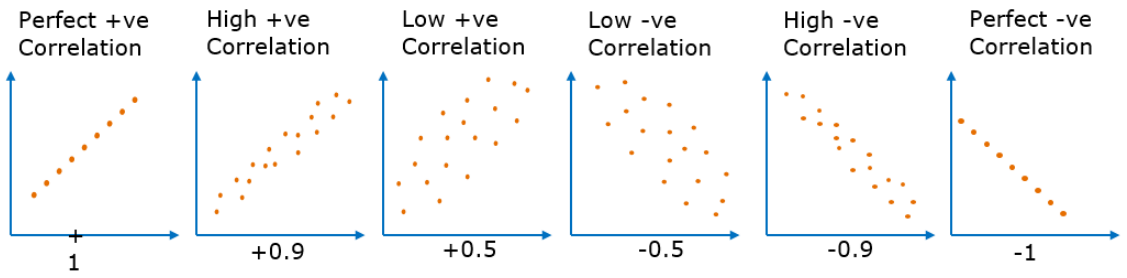


Figure 3.17: Scatterplots with different correlations.

Standardize: $X \to (X-\mu_X)/\sigma_X$, $Y \to (Y-\mu_Y)/\sigma_Y$. Then:

$$\mathbb{E}\left[\frac{X-\mu_X}{\sigma_X}\right] = 0, \qquad \mathrm{Var}\left[\frac{X-\mu_X}{\sigma_X}\right] = 1.$$

After standardization, both variables have mean 0 and variance 1. The correlation is the covariance of the standardized variables.

Definition 3.5.1 (Correlation)**.**

$$\mathrm{Corr}(X,Y) = \mathrm{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

The correlation $\rho = \mathrm{Corr}(X,Y)$ always satisfies $-1 \le \rho \le 1$. It equals $+1$ when Y is a positive linear function of X, -1 when Y is a negative linear function of X, and 0 when there is no linear relationship.

In terms of data:

$$\mathrm{Corr}(X,Y) \approx \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i-\bar{Y})^2}}.$$

3.5.1 Geometric Interpretation: Cosine of an Angle

Center the data: $\tilde{X}_i = X_i - \bar{X}$, $\tilde{Y}_i = Y_i - \bar{Y}$. Then:

$$\text{Corr}(X,Y) \approx \frac{\sum_{i=1}^n \tilde{X}_i \tilde{Y}_i}{\sqrt{\sum_{i=1}^n \tilde{X}_i^2}\sqrt{\sum_{i=1}^n \tilde{Y}_i^2}} = \frac{\langle \tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \rangle}{\|\tilde{\mathbf{X}}\|\,\|\tilde{\mathbf{Y}}\|} = \cos\theta,$$

where θ is the angle between the centered data vectors in $\mathbb{R}^n$. The key identities are:

$$\frac{1}{n}\langle \tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \rangle = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{Y}_i = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \approx \text{Cov}(X,Y),$$

$$\frac{1}{n}\|\tilde{\mathbf{X}}\|^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \approx \text{Var}(X),$$

$$\frac{1}{n}\|\tilde{\mathbf{Y}}\|^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 \approx \text{Var}(Y).$$

Correlation is the cosine of the angle between two centered data vectors. This beautiful geometric interpretation connects statistics to linear algebra. When the vectors point in the same direction ($\theta = 0$), $\cos\theta = 1$. When they are perpendicular ($\theta = 90$), $\cos\theta = 0$. When they point in opposite directions ($\theta = 180$), $\cos\theta = -1$.

3.5.2 Correlation and Regression

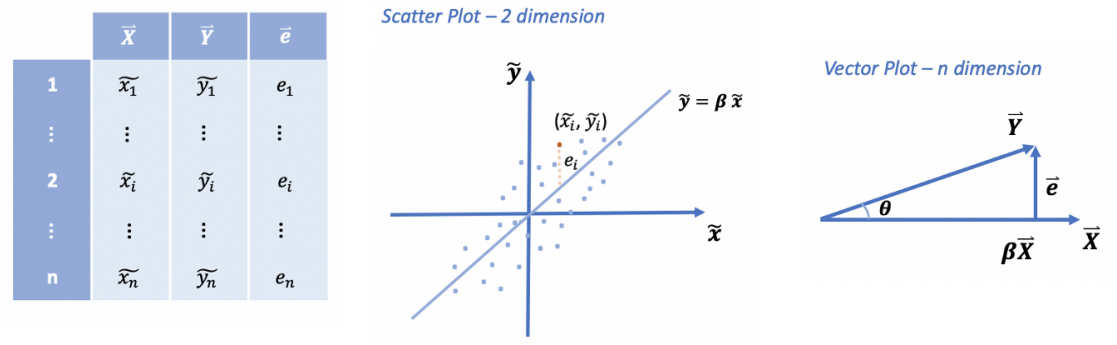


Figure 3.18: Regression: decomposing $\tilde{\mathbf{Y}}$ into a component along $\tilde{\mathbf{X}}$ and a residual $\mathbf{e}$.

Project $\tilde{\mathbf{Y}}$ onto $\tilde{\mathbf{X}}$: this decomposes $\tilde{\mathbf{Y}}$ into a part explained by $\tilde{\mathbf{X}}$ and a residual $\mathbf{e}$ that is perpendicular to $\tilde{\mathbf{X}}$. The strength of the linear relationship is measured by:

$$\frac{\|\mathbf{e}\|^2}{\|\tilde{\mathbf{Y}}\|^2} = \frac{\sum_i e_i^2}{\sum_i (Y_i - \bar{Y})^2} = \sin^2\theta = 1 - \cos^2\theta = 1 - \rho^2.$$

$$\frac{\|\beta\tilde{\mathbf{X}}\|}{\|\tilde{\mathbf{Y}}\|} = \cos\theta = \rho, \qquad \beta = \rho\frac{\|\tilde{\mathbf{Y}}\|}{\|\tilde{\mathbf{X}}\|} = \rho\frac{\sigma_Y}{\sigma_X}.$$

For the bivariate normal with $\sigma_X = \sigma_Y = 1$: $\beta = \rho$ and $1 - \rho^2$ is the fraction of variance unexplained. So ρ^2 tells us what fraction of the variance in Y can be explained by a linear relationship with X.

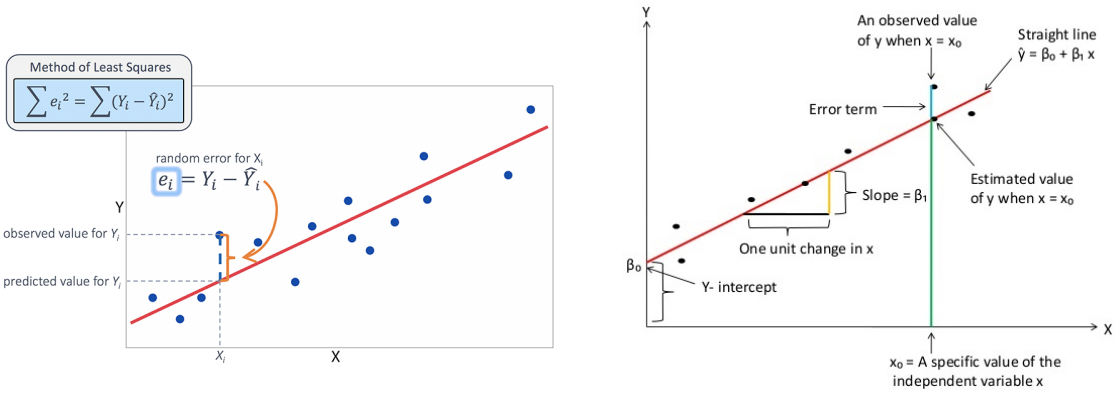


Figure 3.19: Regression line and the explained/unexplained variance.

The regression line is:

$$\hat{Y} - \bar{Y} = \beta_1(X - \bar{X}), \qquad \hat{Y} = \beta_1 X + \beta_0.$$

Multiple regression extends this to several predictors:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

3.5.3 Deep Learning: Nonlinear Regression

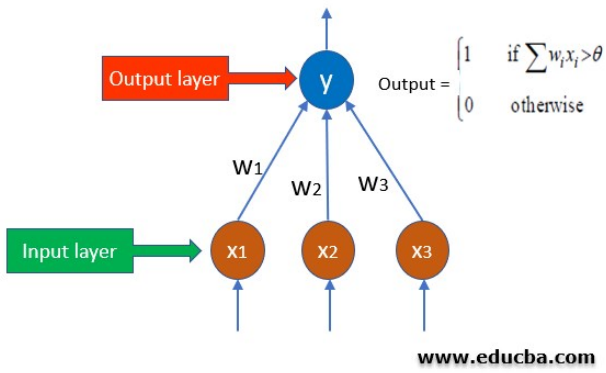


Figure 3.20: A perceptron: the building block of neural networks. Rectified Linear Unit: $\text{ReLU}(a) = \max(0, a)$.

Linear regression can only capture linear relationships. For nonlinear relationships, we use neural networks. A single perceptron computes:

$$y = \max\left(0, \sum_i w_i x_i + b\right).$$

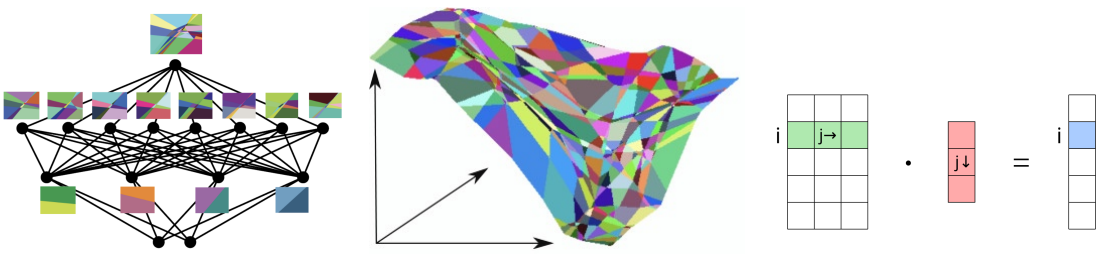


Figure 3.21: Multi-layer perceptrons: each node applies linear combination then ReLU.

Each node: linear combination of nodes at the layer below $\sum_i w_i x_i$, then ReLU $\max(0, \sum_i w_i x_i - \theta)$. A multi-layer perceptron stacks layers:

$$h_l = \max(0, W_l h_{l-1} + b_l).$$

h_l: embedding, encoding, representation, thought vector. W_l: weight matrix. b_l: bias vector. The result is a piecewise linear mapping from input to output. Weights can be learned from training data and then used for testing.

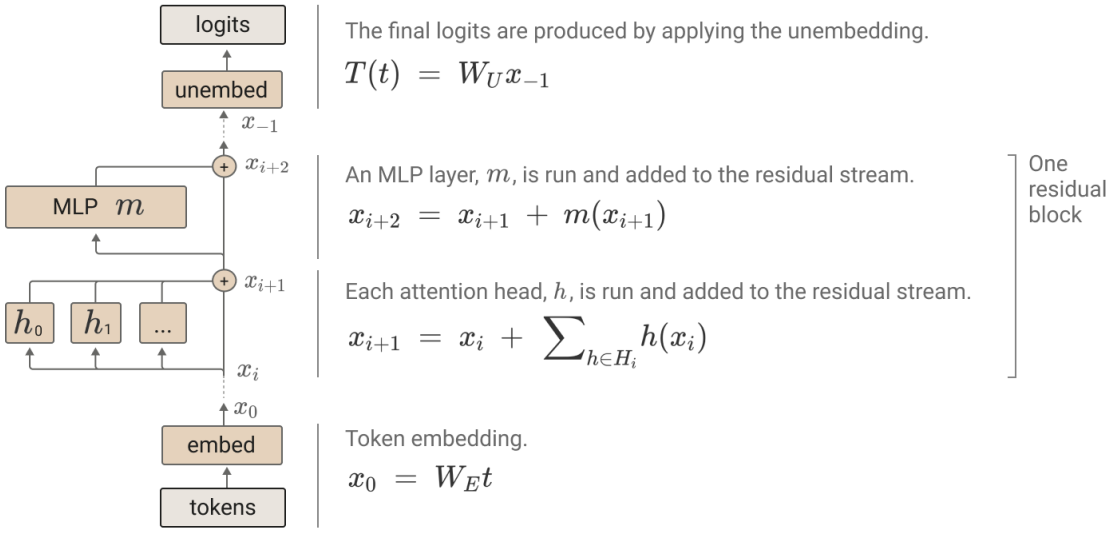


Figure 3.22: The GPT architecture. Embed: word → vector. Compute: vectors operated by learned matrices. Unembed: vector → probabilities for next word.

3.6 Independence and the Variance of Sums

3.6.1 Independence Implies Zero Covariance

Recall the definitions: $P(A \cap B) = P(A)P(B)$; $p(x, y) = p_X(x)\, p_Y(y)$; $p(y \mid x) = p_Y(y)$; $f(x, y) = f_X(x)\, f_Y(y)$; $f(y \mid x) = f_Y(y)$.

If X and Y are independent, we show step by step that $\text{Cov}(X, Y) = 0$:

$$\begin{aligned}
\text{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)\, p(x,y) \\
&= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)\, p_X(x)\, p_Y(y) \qquad \text{(independence)} \\
&= \underbrace{\left(\sum_x (x-\mu_X)\, p_X(x)\right)}_{=\mathbb{E}(X)-\mu_X=0} \underbrace{\left(\sum_y (y-\mu_Y)\, p_Y(y)\right)}_{=\mathbb{E}(Y)-\mu_Y=0} = 0.
\end{aligned}$$

The key step is using independence to factor $p(x, y)$ into $p_X(x)\, p_Y(y)$, which allows the double sum to factor into a product of two single sums. Each single sum equals zero because it is the mean deviation from the mean.

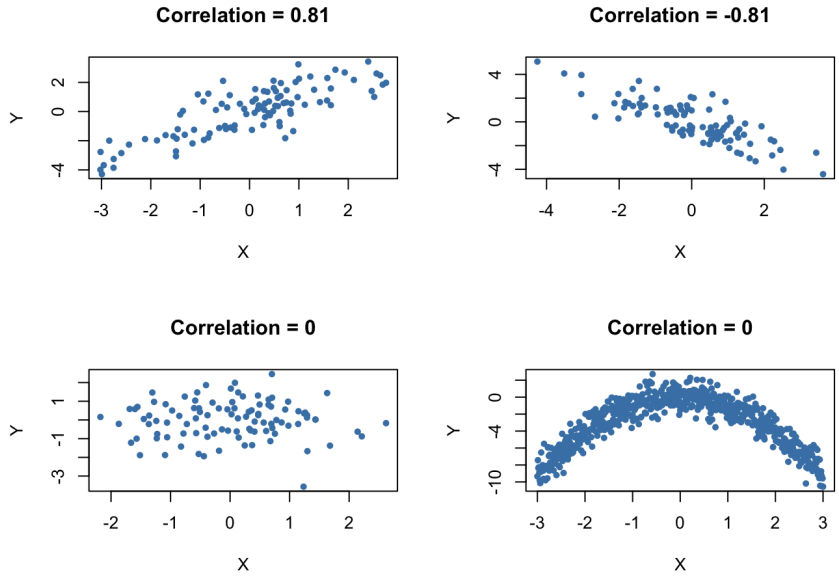


Figure 3.23: Correlation patterns in bivariate data.

Remark 3.6.1. The converse is *not* true: zero covariance does not imply independence.

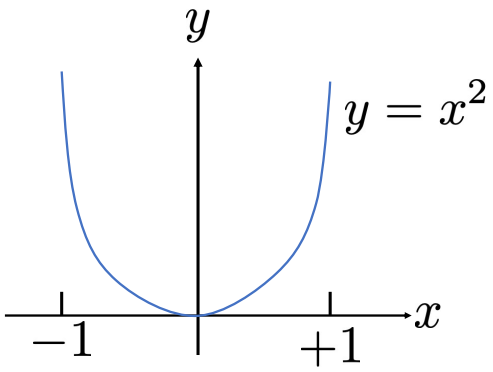


Figure 3.24: X uniform on $[-1, 1]$ and $Y = X^2$: dependent but uncorrelated.

Let X be uniform over $[-1, 1]$, $Y = X^2$. Then X and Y are clearly not independent — Y is completely determined by X! However, $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0$ (by symmetry, since X^3 is an odd function of a symmetric random variable), and $\mathbb{E}(X) = 0$. Thus $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

The lesson: covariance and correlation only measure *linear* relationships. X and $Y = X^2$ have a strong nonlinear relationship, but their covariance is zero.

For the bivariate normal, however, uncorrelated *does* imply independent (a special property of the normal distribution).

3.6.2 Bivariate Normal: Covariance Equals ρ

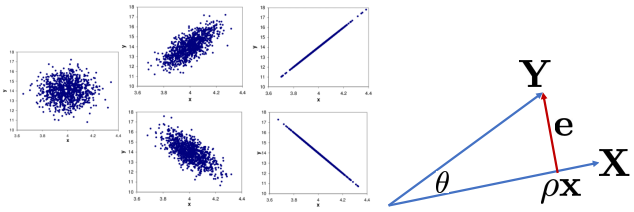


Figure 3.25: Bivariate normal: $Y = \rho X + \epsilon$ with ϵ independent of X.

For the bivariate normal $Y = \rho X + \epsilon$ with $X \sim \mathcal{N}(0, 1)$ and $\epsilon \sim \mathcal{N}(0, 1 - \rho^2)$ independent of X:

$$\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(\rho X + \epsilon) = 0. \\
\mathrm{Var}(Y) &= \mathrm{Var}(\rho X + \epsilon) = \rho^2 \,\mathrm{Var}(X) + \mathrm{Var}(\epsilon) = \rho^2 + (1 - \rho^2) = 1. \\
\mathrm{Cov}(X, Y) &= \mathrm{Cov}(X, \rho X + \epsilon) = \rho \,\mathrm{Cov}(X, X) + \mathrm{Cov}(X, \epsilon) = \rho \cdot 1 + 0 = \rho.
\end{aligned}$$

The last line uses the linearity of covariance and the fact that $\mathrm{Cov}(X, \epsilon) = 0$ (since ϵ is independent of X). So the parameter ρ in the bivariate normal construction is precisely the correlation between X and Y.

3.6.3 Variance of a Sum

For any two random variables (not necessarily independent):

$$\mathbb{E}(X + Y) = \sum_x \sum_y (x + y)\, p(x, y) = \sum_x \sum_y x\, p(x, y) + \sum_x \sum_y y\, p(x, y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

The expectation of a sum is always the sum of expectations — regardless of dependence.

The variance of a sum is more subtle. We derive it step by step:

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[((X+Y)-\mu_{X+Y})^2] \\ &= \mathbb{E}[((X-\mu_X)+(Y-\mu_Y))^2] \\ &= \mathbb{E}[(X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y)] \\ &= \mathbb{E}[(X-\mu_X)^2]+\mathbb{E}[(Y-\mu_Y)^2]+2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y).\end{aligned}$$

The cross term $2\mathrm{Cov}(X,Y)$ captures the interaction between X and Y.

If X and Y are independent (so $\mathrm{Cov}(X,Y)=0$):

$$\boxed{\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).}$$

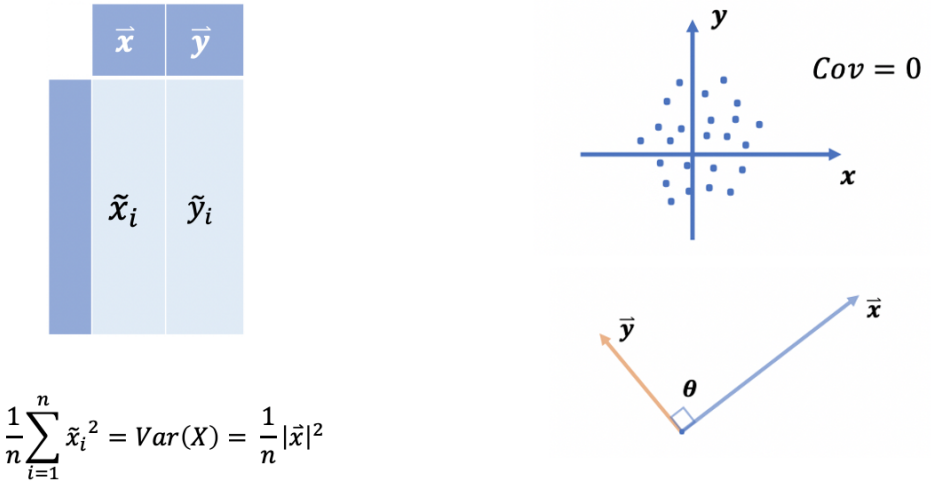


Figure 3.26: The variance of a sum depends on the covariance structure.

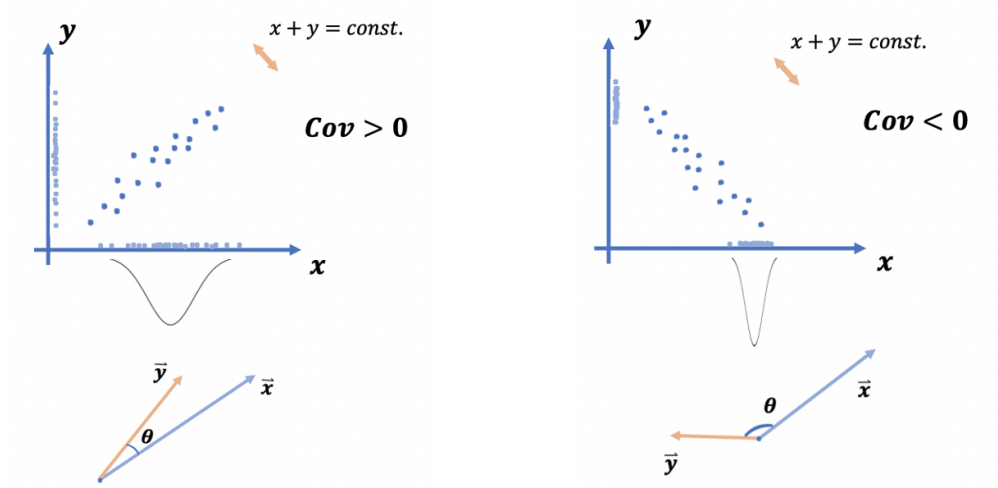


Figure 3.27: More examples of variance of sums.

If X and Y are positively correlated, $\text{Var}(X+Y) > \text{Var}(X) + \text{Var}(Y)$ — the variability is amplified because X and Y tend to be large (or small) at the same time. If they are negatively correlated, $\text{Var}(X+Y) < \text{Var}(X)+\text{Var}(Y)$ — the variability is dampened because when X is large, Y tends to be small, and vice versa. This is the mathematical basis of diversification in investing.

3.7 The Law of Large Numbers

3.7.1 Average of IID Random Variables

Let $X_1, X_2, \ldots, X_n$ be **independent and identically distributed (iid)** from a distribution with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$.

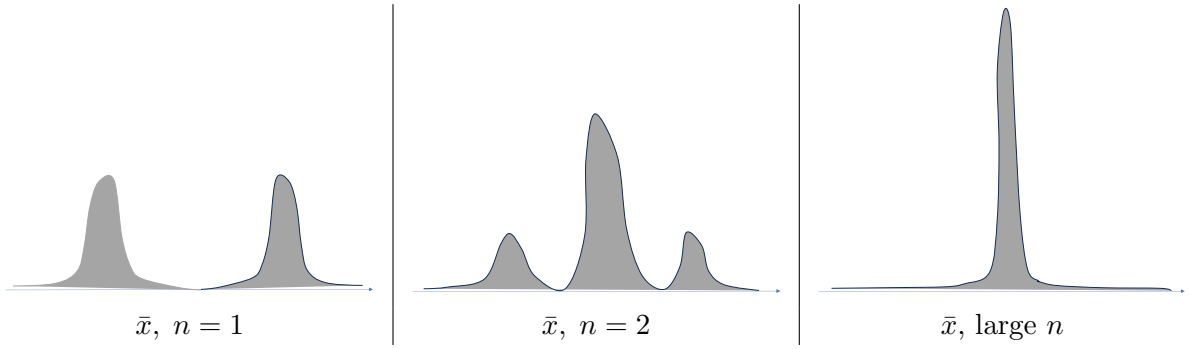


Figure 3.28: As n increases, the distribution of $\bar{X}$ concentrates around μ. For $n = 2$: small + small = small; small + large = medium; large + small = medium; large + large = large. Variance becomes smaller, distribution becomes smoother.

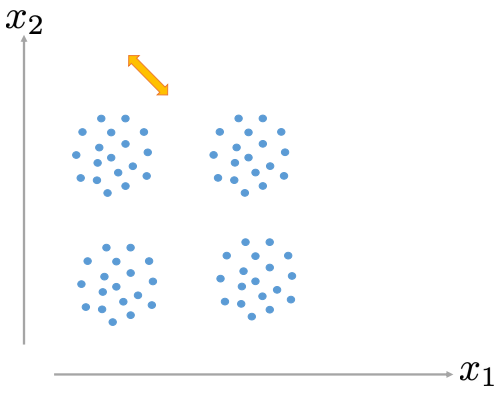


Figure 3.29: The averaging effect: extreme values cancel out.

Define $S = \sum_{i=1}^{n} X_i$ and $\bar{X} = S/n$. Using the linearity of expectation and the additivity of variance for independent random variables:

$$\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mu.$$

$$\mathrm{Var}(S) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\sigma^2.$$

$$\mathbb{E}(\bar{X}) = \frac{\mathbb{E}(S)}{n} = \mu.$$

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The mean of $\bar{X}$ is μ regardless of n, but the variance shrinks like $1/n$. As n increases, the distribution of $\bar{X}$ becomes more and more concentrated around μ.

Theorem 3.7.1 (Law of large numbers)**.** $\mathrm{Var}(\bar{X}) = \sigma^2/n \to 0$ *as* $n \to \infty$. *Therefore* $\bar{X} \to \mu$ *in probability:*

$$P(|\bar{X} - \mu| < \epsilon) \to 1 \qquad \textit{for any } \epsilon > 0.$$

Average $\to$ *expectation.*

The proof is elegant in its simplicity. By Chebyshev's inequality, $P(|\bar{X} - \mu| \geq \epsilon) \leq \mathrm{Var}(\bar{X})/\epsilon^2 = \sigma^2/(n\epsilon^2) \to 0$. As the variance shrinks to zero, the probability of being far from the mean goes to zero.

3.7.2 Special Case: Coin Flipping

For $Z_i \sim \text{Bernoulli}(p)$ iid, $X = \sum Z_i \sim \text{Binomial}(n, p)$:

$$\mathbb{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p), \quad \mathbb{E}(X/n) = p, \quad \mathrm{Var}(X/n) = p(1-p)/n \to 0.$$

So $X/n \to p$: frequency $\to$ probability. This is the precise statement of the intuition from Chapter 1.

X/n is the average of Z_i. Probability p is the expectation of Z_i.

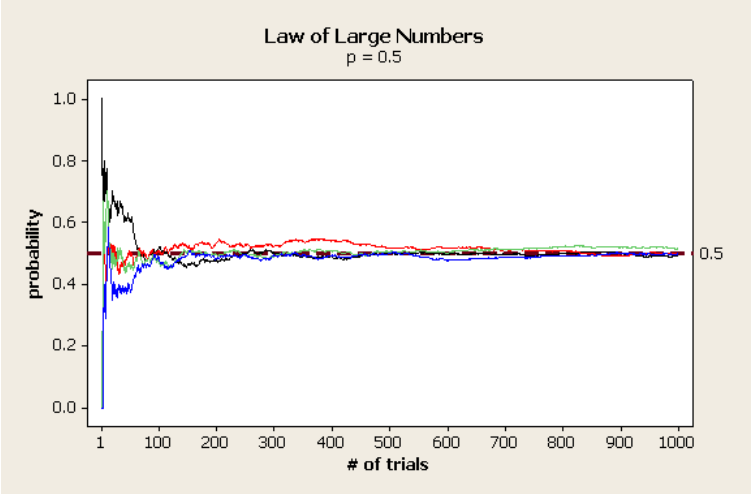


Figure 3.30: Frequency of heads converging to 1/2. Intuition: most of 2^n sequences have frequencies close to 1/2.

3.7.3 N^n Reasoning

Ω_1: population of N people. Each person $a \in \Omega_1$ has $X(a) =$ height. Population average $\mu = \mathbb{E}(X)$.

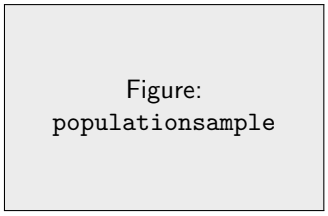


Figure 3.31: N^n reasoning: repeat random sampling n times independently.

$\to N^n$ equally likely sequences: Ω_n. For a sequence $\omega \in \Omega_n$, $\bar{X}(\omega) =$ sequence average. $A = \{\omega : |\bar{X}(\omega) - \mu| \leq 0.01\}$: representative sequences. $P(A) = |A|/|\Omega_n| \to 1$ as $n \to \infty$.

3.7.4 Concentration of Measure in the Cube

Special case: $X_i \sim \text{Uniform}[0,1] = \Omega_1$, iid, $i = 1, \ldots, n$.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \to \mathbb{E}(X_i) = 1/2, \qquad P(|\bar{X} - 1/2| < 0.01) \to 1.$$

Intuition: A sequence $(X_1, \ldots, X_n)$ is a random point in $\Omega_n = [0,1]^n$, the n-dimensional unit cube. The set $A = \{(x_1, \ldots, x_n) : |\bar{x} - 1/2| < 0.01\}$ is the central diagonal piece. $P(A)$ is the volume of A, and $P(A) \to 1$.

The volume of the central diagonal piece is almost the same as the volume of the whole n-dimensional cube. **Most of the points in Ω belong to A.** This is the **concentration of measure** phenomenon — one of the most remarkable facts about high-dimensional geometry.

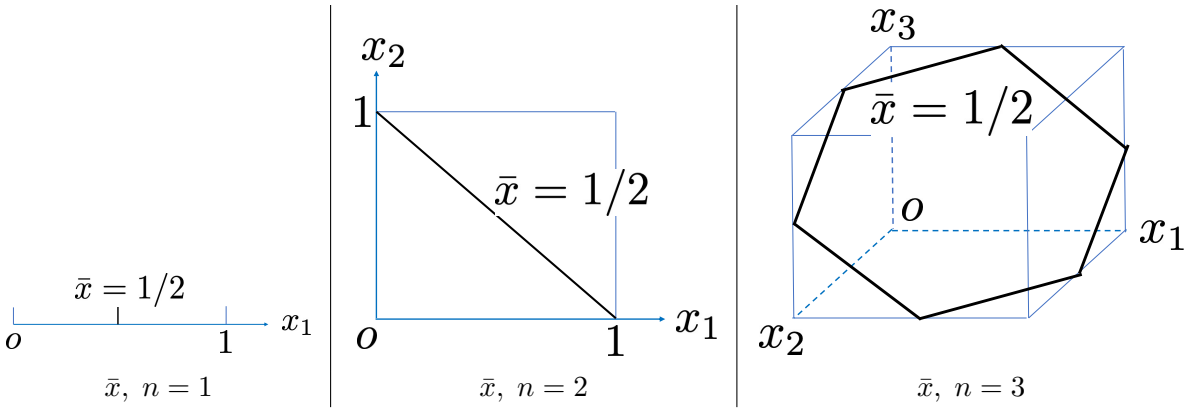


Figure 3.32: Concentration of measure: as n increases, most of the volume concentrates near $\bar{x} = 1/2$.

Remark 3.7.1 (Statistical physics)**.** Suppose $(x_1, \ldots, x_n)$ describes a physical system with $n = 10^{23}$ molecules. The system evolves **deterministically** over time, by traversing Ω. If the system is **ergodic** (it traverses every point in Ω with equal number of visits in the long run), then at any **random moment**, $(x_1, \ldots, x_n) \sim \text{Unif}(\Omega)$. Then most likely it will be in A, with fixed statistical properties (temperature, pressure, magnetism). This is why macroscopic properties of gases are stable even though individual molecules are moving chaotically.

3.8 The Central Limit Theorem

The law of large numbers says $\bar{X} \to \mu$. But it does not tell us the *shape* of the distribution of $\bar{X}$ before the limit. The **central limit theorem** fills this gap: it tells us that the distribution of $\bar{X}$ is approximately normal, regardless of the original distribution of the X_i.

3.8.1 Statement

Let $X_i \sim f(x)$ iid with $\mathbb{E}(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$. Define $S = \sum_{i=1}^n X_i$ and $\bar{X} = S/n$. Then:

$$\mathbb{E}(S) = n\mu, \quad \text{Var}(S) = n\sigma^2, \quad \mathbb{E}(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \sigma^2/n.$$

The deviation $\bar{X} - \mu \to 0$. But if we magnify by $\sqrt{n}$:

$$Y_n = \sqrt{n}(\bar{X} - \mu), \qquad \mathbb{E}(Y_n) = 0, \qquad \text{Var}(Y_n) = (\sqrt{n})^2 \frac{\sigma^2}{n} = \sigma^2.$$

Normalization = (random variable − mean) / standard deviation:

$$Z_n = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \frac{S - n\mu}{\sqrt{n}\,\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Theorem 3.8.1 (Central limit theorem)**.** $Z_n \to \mathcal{N}(0, 1)$ *in distribution:*

$$P(Z_n \in [a, b]) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz,$$

regardless of the original distribution of X_i, discrete or continuous.

This is one of the most remarkable theorems in all of mathematics. No matter what distribution the individual X_i come from — uniform, exponential, Bernoulli, or any other — the normalized sum always converges to the same bell-shaped curve. This universality is why the normal distribution appears so often in nature: many real-world quantities are sums (or averages) of many small independent contributions.

3.8.2 Coin Flipping and Random Walk

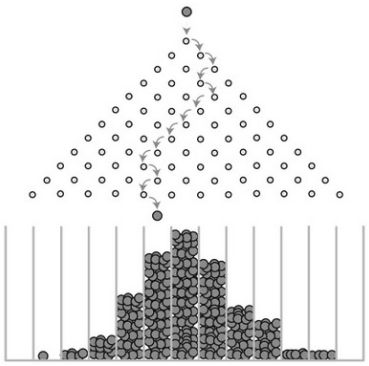


Figure 3.33: The Galton board realizes the CLT.

For $\epsilon_i \sim \text{Bernoulli}(1/2)$ iid, $X = \sum \epsilon_i \sim \text{Binomial}(n, 1/2)$:

$$\mu = \mathbb{E}(X) = n/2, \quad \sigma^2 = \text{Var}(X) = n/4.$$

$$P\left(Z = \frac{X - n/2}{\sqrt{n}/2} = z\right) \approx \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \cdot \frac{2}{\sqrt{n}} = f(z)\,\Delta z.$$

In general, ϵ_i can be any discrete or continuous random variable with $\mathbb{E}(\epsilon_i) = 0$.

3.8.3 Die Rolling

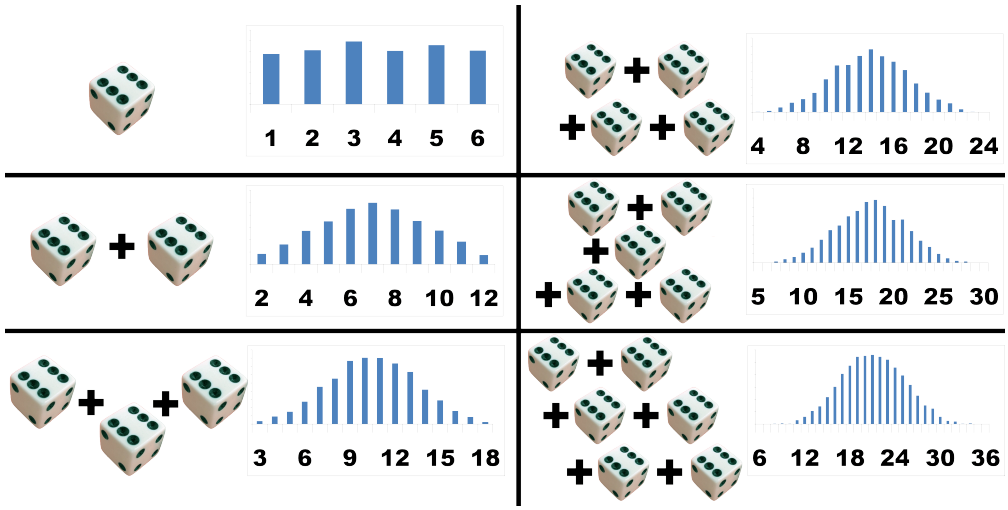


Figure 3.34: Rolling a die repeatedly: the distribution of $S = \sum X_i$ becomes bell-shaped.

Repeat rolling a die and plot the histogram of the sum $S = \sum_{i=1}^{n} X_i$. Even though each die roll has a flat (uniform) distribution, the sum quickly takes on a bell shape. Then $S \approx \mathcal{N}(n\mu, n\sigma^2)$ and $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$.

3.8.4 Population of Sequences

6^n equally likely sequences $\rightarrow$ 6^n equally likely sums $\rightarrow$ bell-shaped histogram.

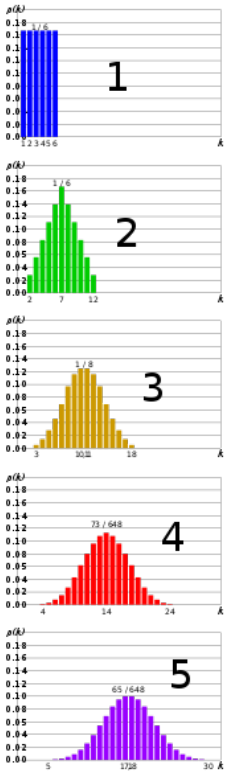


Figure 3.35: The 6^n sums form a bell curve.

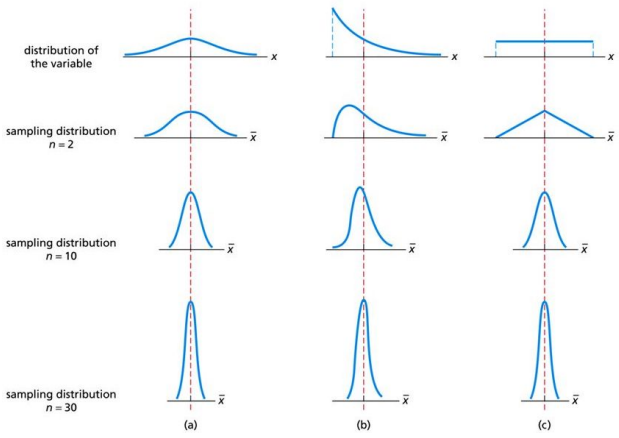


Figure 3.36: Central limit theorem: the histogram of the normalized sum converges to the normal density.

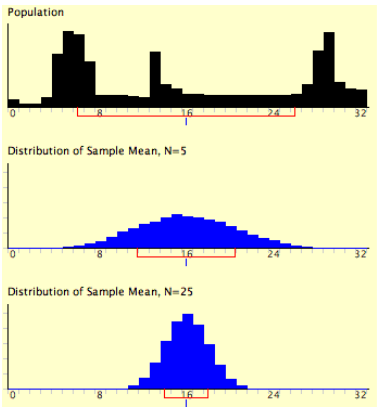


Figure 3.37: Universality: regardless of the distribution of X_i, the normalized sum $Z_n \to \mathcal{N}(0,1)$.

3.9 Take-Home Messages

As long as you can count (and average):

(1) **Population of equally likely possibilities.** Probability = population proportion.

(2) **Large sample of repetitions.** Frequency (fluctuating) $\approx$ probability (fixed).

(3) N^n **reasoning:** hyper-population of sequences. (1) $\to$ (2).

The three key quantities: (a) **Probability**: population proportion, long-run frequency. (b) **Expectation**: population average, long-run average. (c) **Conditional**: sub-population, when something happens.

Forward conditional: cause $\to$ effect. **Backward** conditional: effect $\to$ cause. Population migration: cause state $\to$ effect state. **Continuous**: discretize, infinitesimal analysis.

Chapter 4

Advanced Topics

This chapter brings together several advanced themes: continuous-time stochastic processes, proofs of normal approximation, conditional independence and graphical models, transformations of random variables, simulation methods, convexity and the Jensen inequality, and information theory. The unifying thread is the same as throughout the book: we discretize time and space, count equally likely possibilities, and take limits.

Each topic in this chapter follows the same recipe. We start with a discrete, finite, countable version of the problem — something we can understand by counting. Then we take a limit (usually $\Delta t \to 0$ or $n \to \infty$) to obtain a continuous result. Along the way, we will see how the exponential function, the Poisson distribution, Brownian motion, and the normal distribution all emerge naturally from this discrete-to-continuous passage.

4.1 From Discrete to Continuous Time

Many real-world processes evolve in *continuous* time. To understand them, we start with discrete time and take limits. The key idea is to think of continuous time as the limit of many tiny discrete time steps — just like a movie is really a sequence of still images shown rapidly enough to create the illusion of continuous motion.

4.1.1 Particle Decay

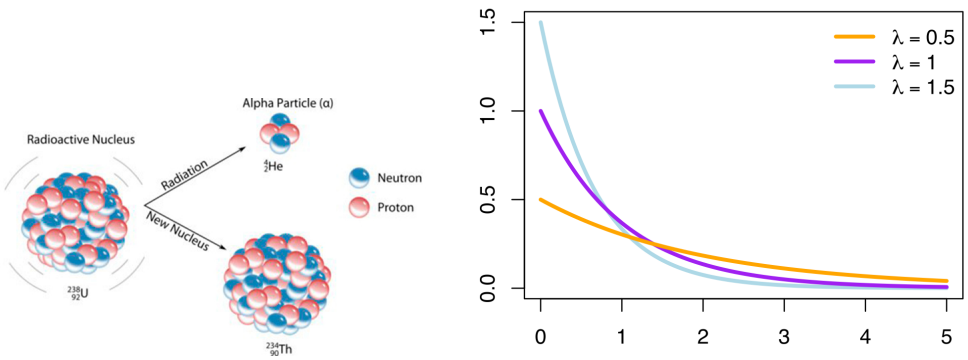


Figure 4.1: Left: a decaying particle. Right: the exponential density.

T: time until decay. $T \sim \text{Exponential}(\lambda)$. $P(T \in (t, t+\Delta t)) = f(t)\,\Delta t = \lambda e^{-\lambda t}\,\Delta t$.

We studied the exponential distribution in Chapter 2 as an abstract density function. Now we will see *where it comes from* — it arises naturally as the continuous-time limit of a sequence of coin flips.

4.1.2 Making a Movie

Figure 4.2: Divide time into small intervals of length Δt, like frames in a movie.

Divide the time into small intervals of length Δt (e.g., 1/24 second, or 1/100 second). Show a "picture" at times $0, \Delta t, 2\Delta t, \ldots$ This gives an illusion of a continuous-time process as $\Delta t \to 0$.

This "making a movie" technique is one of the most powerful tools in applied mathematics. Whenever we face a continuous-time process, we discretize time into tiny intervals, analyze the discrete version using the counting tools from Chapters 1–3, and then take the limit $\Delta t \to 0$. The continuous-time result inherits all the structure of the discrete version.

4.1.3 The Exponential Function from Compound Interest

Before we can derive continuous-time distributions, we need to understand how the exponential function e^x arises. The most intuitive route is through compound interest.

Divide $[0, t]$ into n small intervals, $\Delta t = t/n$. With interest rate r:

Starting with 1 dollar at time 0:

$$\begin{aligned}
&\text{Time } 0: \quad 1.\\
&\text{Time } \Delta t: \quad (1 + r\Delta t).\\
&\text{Time } 2\Delta t: \quad (1 + r\Delta t)^2.\\
&\text{Time } 3\Delta t: \quad (1 + r\Delta t)^3.\\
&\quad \vdots\\
&\text{Time } t = n\Delta t: \quad (1 + r\Delta t)^n.
\end{aligned}$$

Let us make sure each line is clear. At time Δt, you have your original dollar plus interest earned during the first interval: $1 + r\Delta t$. At time $2\Delta t$, you earn interest on the *new* balance (this is "compound" interest): $(1 + r\Delta t)(1 + r\Delta t) = (1 + r\Delta t)^2$. Each interval multiplies your balance by the factor $(1 + r\Delta t)$, so after n intervals, your balance is $(1 + r\Delta t)^n$.

As $n \to \infty$ (or $\Delta t \to 0$):

$$\left(1 + r\frac{t}{n}\right)^n \to e^{rt}.$$

This follows from the fundamental limit $\left(1 + \frac{1}{n}\right)^n \to e$, which gives the approximation:

$$1 + \Delta x \approx e^{\Delta x} \qquad \text{for small } \Delta x.$$

Therefore:

$$(1 + r\Delta t)^{t/\Delta t} \approx (e^{r\Delta t})^{t/\Delta t} = e^{rt}.$$

This approximation $1 + \Delta x \approx e^{\Delta x}$ (valid for small Δx) will be used repeatedly throughout this chapter. It is the bridge between discrete multiplicative processes and continuous exponential growth or decay.

4.2 The Poisson Process

The Poisson process models events that occur randomly in time: radioactive decays, phone calls, earthquakes, typos. It is one of the most widely used models in science and engineering, and it arises naturally from the "coin flip in each small interval" construction.

4.2.1 Construction: Coin Flips in Small Intervals

Figure 4.3: Flip a coin within each small time interval.

Divide time into small intervals of length Δt. Within each interval, flip a coin: with probability $p = \lambda \Delta t$, an event occurs; with probability $1 - \lambda \Delta t$, nothing happens. Here λ is the **rate** or **intensity**.

The key assumption is that the probability of an event in a small interval is proportional to the length of that interval. If you make the intervals twice as short, the probability in each interval is half as large. The proportionality constant λ tells us the average number of events per unit time.

For example, $\Delta t = 1$ hour, the event happens once every 10 years on average, so $\lambda \Delta t = 1/(3650 \times 24)$.

4.2.2 Geometric $\to$ Exponential

The waiting time T until the first event has a geometric distribution (in units of Δt). We compute the probability step by step:

$$\begin{aligned} P(T \in (t, t+\Delta t)) &= (1-\lambda\Delta t)^{t/\Delta t} \cdot \lambda \Delta t \\ &\approx (e^{-\lambda \Delta t})^{t/\Delta t} \cdot \lambda \Delta t \qquad (\text{using } 1-\lambda\Delta t \approx e^{-\lambda \Delta t}) \\ &= e^{-\lambda t}\, \lambda\, \Delta t. \end{aligned}$$

Let us unpack this calculation. The event "T falls in the interval $(t, t+\Delta t)$" means: nothing happens in the first $t/\Delta t$ intervals (each with probability $1-\lambda\Delta t$), and then an event occurs in the next interval (with probability $\lambda \Delta t$). By independence, we multiply all these probabilities together. The first $t/\Delta t$ factors give $(1-\lambda\Delta t)^{t/\Delta t}$, which we approximate by $e^{-\lambda t}$ using our small-Δx formula.

Dividing by Δt gives the density: $f(t) = \lambda e^{-\lambda t}$, the **exponential distribution**.

The survival probability (no event by time t) is:

$$P(T > t) = (1-\lambda\Delta t)^{t/\Delta t} \approx (e^{-\lambda\Delta t})^{t/\Delta t} = e^{-\lambda t}.$$

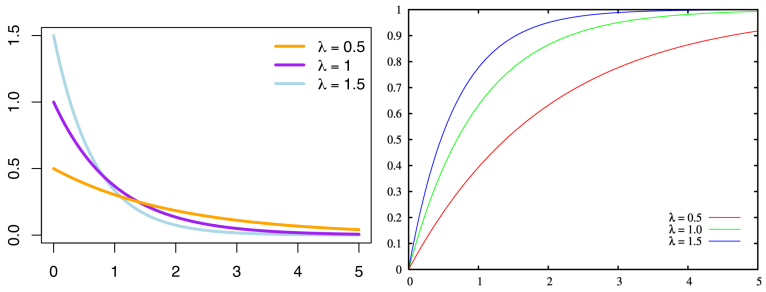


Figure 4.4: Exponential density and CDF.

We can write $T = \tilde{T}\,\Delta t$ where $\tilde{T} \sim \text{Geometric}(p = \lambda\Delta t)$. Then:

$$\mathbb{E}(T) = \mathbb{E}(\tilde{T})\,\Delta t = \frac{1}{p}\,\Delta t = \frac{1}{\lambda\Delta t}\,\Delta t = \frac{1}{\lambda}.$$

This confirms what we derived in Chapter 2: $\mathbb{E}(T) = 1/\lambda$. But now we see *why* this formula holds — it comes directly from the geometric expectation $1/p$, scaled by the time unit Δt.

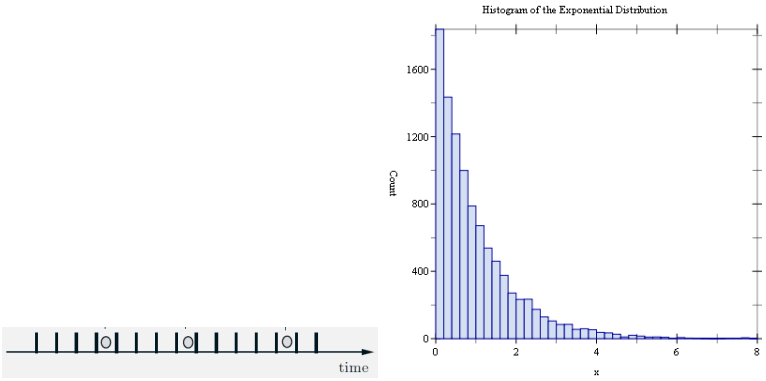


Figure 4.5: 1 million particles decay at different times. Each small period is a bin.

4.2.3 Binomial → Poisson

Now we count the *number* of events in a fixed time interval. Let X be the number of events in $[0, t]$. Since we flip a coin $n = t/\Delta t$ times independently, each with probability $p = \lambda\Delta t$:

$$X \sim \text{Binomial}(n, p), \qquad \mathbb{E}(X) = np = (t/\Delta t)(\lambda\Delta t) = \lambda t.$$

$\lambda = \mathbb{E}(X)/t$: rate or intensity (expected number of events per unit time).

As $\Delta t \to 0$ (equivalently $n \to \infty$ with $np = \lambda t$ fixed):

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \to \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

This is a beautiful limit: as we make the time intervals smaller and smaller (increasing n) while keeping the overall rate constant (so p shrinks proportionally), the binomial distribution converges to a new distribution — the Poisson distribution. The Poisson distribution is the natural model for "rare events": each individual interval has a tiny probability of producing an event, but there are very many intervals.

4.2.4 Derivation of the Poisson Limit

The derivation proceeds by careful asymptotic analysis, step by step:

$$\begin{aligned} P(X=k) &= \frac{n(n-1)\cdots(n-k+1)}{k!}\, p^k\,(1-p)^{n-k} \\ &= \frac{(t/\Delta t)(t/\Delta t-1)\cdots(t/\Delta t-k+1)}{k!}\,(\lambda\Delta t)^k\,(1-\lambda\Delta t)^{t/\Delta t-k} \\ &= \frac{t(t-\Delta t)(t-2\Delta t)\cdots(t-(k-1)\Delta t)}{k!}\,\lambda^k\,(1-\lambda\Delta t)^{t/\Delta t}\,(1-\lambda\Delta t)^{-k}. \end{aligned}$$

In the second line, we substitute $n = t/\Delta t$ and $p = \lambda\Delta t$. In the third line, we multiply through: $n \cdot \lambda\Delta t = t/\Delta t \cdot \lambda\Delta t = \lambda t$. More carefully, we separate $(1-\lambda\Delta t)^{n-k}$ into $(1-\lambda\Delta t)^n \cdot (1-\lambda\Delta t)^{-k}$.

As $\Delta t \to 0$: the numerator $t(t-\Delta t)\cdots \to t^k$ because each factor approaches t; $(1-\lambda\Delta t)^{t/\Delta t} \to e^{-\lambda t}$ by our compound interest formula; $(1-\lambda\Delta t)^{-k} \to 1$ because k is fixed while $\Delta t \to 0$. Therefore:

$$P(X=k) \to \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}.$$

Definition 4.2.1 (Poisson distribution)**.** $X \sim \text{Poisson}(\lambda t)$ has pmf $P(X=k) = \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}$ for $k = 0, 1, 2, \ldots$, with $\mathbb{E}(X) = \lambda t$.

Let us verify that the probabilities sum to 1: $\sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!}\, e^{-\lambda t} = e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\lambda t} \cdot e^{\lambda t} = 1$, using the Taylor series $e^x = \sum_{k=0}^{\infty} x^k/k!$.

4.3 Brownian Motion and Diffusion

Brownian motion is one of the most important stochastic processes in mathematics and physics. It models the erratic, jittery motion of particles suspended in a fluid, and it is also the mathematical foundation of modern finance (the Black–Scholes model for option pricing).

4.3.1 Dust Particles in Water

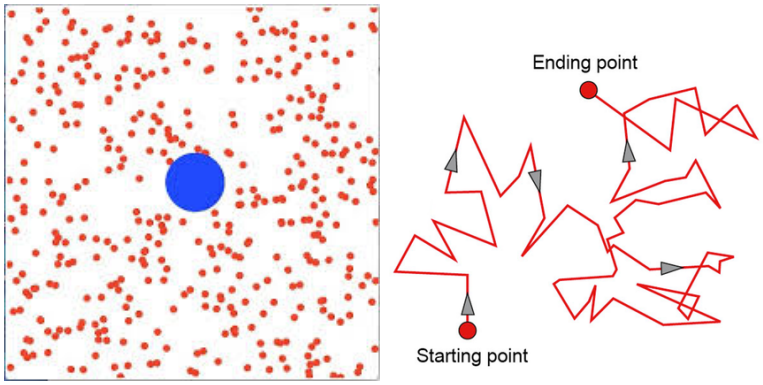


Figure 4.6: Brownian motion: the erratic path of a dust particle in water.

When Robert Brown observed pollen grains in water under a microscope in 1827, he saw them jiggling erratically. This **Brownian motion** is caused by random bombardment of water molecules.

Each water molecule is far too small to see, but their collective, random impacts cause the much larger pollen grain to wander aimlessly. The mathematical model of this phenomenon turns out to be a random walk taken to its continuous-time limit.

4.3.2 Recall: Random Walk

Either go forward or backward by flipping a fair coin.

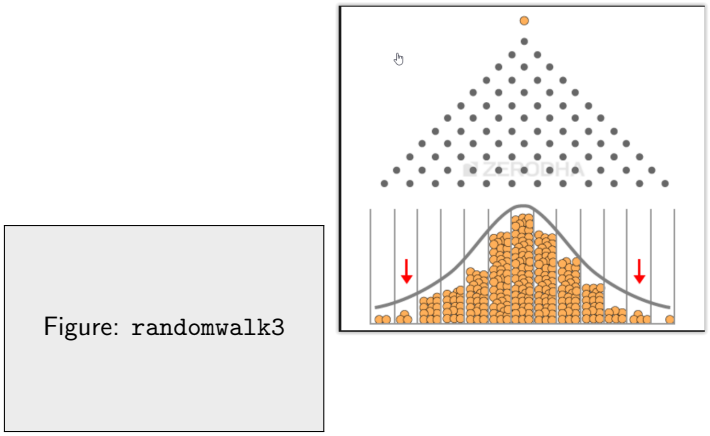


Figure 4.7: Left: discrete random walk. Right: Brownian motion paths.

Number of heads $Y \sim \text{Binomial}(n, 1/2)$, then random walk ends up at $X = Y - (n - Y) = 2Y - n$. Writing $X = \epsilon_1 + \epsilon_2 + \cdots + \epsilon_n$ where $\epsilon_i = +1$ or -1 with probability $1/2$ each.

4.3.3 Discretize Time and Space

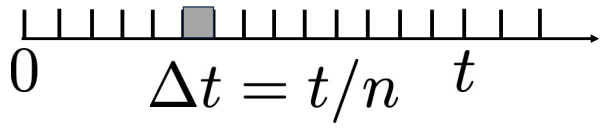


Figure 4.8: Discretize time and space: move $\pm\Delta x$ every Δt seconds.

To go from a random walk to Brownian motion, we need to discretize *both* time and space and then take limits:

(1) **Time:** Divide $[0, t]$ into n intervals, $\Delta t = t/n$ (time unit).

(2) **Space:** Within each small time interval, move forward or backward by Δx (space unit).

$P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$, ϵ_i independent. The position after n steps is:

$$X = \sum_{i=1}^{n} \epsilon_i \, \Delta x = (2Y - n) \, \Delta x.$$

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(\epsilon_i) \, \Delta x = 0.$$

The expected position is zero because each step is equally likely to go left or right — on average, the particle stays where it started.

4.3.4 The Diffusion Scaling

The variance of the position is:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(\epsilon_i)\,(\Delta x)^2 = n\,(\Delta x)^2 = \frac{t}{\Delta t}\,(\Delta x)^2.$$

Here we use $\mathrm{Var}(\epsilon_i) = 1$ (since $\epsilon_i = \pm 1$ with equal probability) and the independence of the ϵ_i (so the variance of the sum is the sum of the variances). The result is n times $(\Delta x)^2$.

For this to remain finite as $\Delta t \to 0$, we need:

$$(\Delta x)^2/\Delta t = \sigma^2 \qquad \text{(a constant)}, \qquad \text{so} \qquad \Delta x = \sigma\sqrt{\Delta t}.$$

Then $\mathrm{Var}(X) = \sigma^2 t$. The velocity is:

$$\text{velocity} = \Delta x/\Delta t = \sigma/\sqrt{\Delta t} \to \infty!$$

The particle moves infinitely fast in the limit. Einstein showed that σ is related to the size of molecules, providing a way to measure atomic scales from macroscopic observations.

The scaling $\Delta x = \sigma\sqrt{\Delta t}$ is the famous **diffusion scaling**. It says that the step size in space must shrink as the *square root* of the time step. If we halve the time step, we only reduce the space step by a factor of $\sqrt{2} \approx 1.41$, not 2. This square-root relationship is the signature of diffusion.

4.3.5 Brownian Motion

Figure: `randomwalk3`

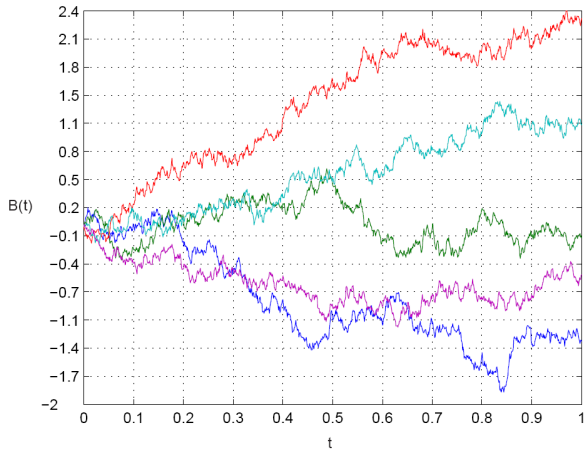


Figure 4.9: Left: discrete random walk. Right: continuous Brownian motion trajectories.

In the limit, we obtain **Brownian motion**:

$$X_{t+\Delta t} = X_t + \sigma\sqrt{\Delta t}\,\epsilon_t,$$

where $\mathbb{E}(\epsilon_t) = 0$, $\mathrm{Var}(\epsilon_t) = 1$, and the ϵ_t are iid. The path is **nowhere differentiable**.

The nowhere-differentiability is a consequence of the infinite velocity we computed above. At every instant, the particle is changing direction so rapidly that there is no well-defined velocity. The path is continuous (no jumps) but infinitely jagged — it is a fractal.

By the central limit theorem (sum of independent random variables $\to$ normal), the position at time t is:

$$X \sim \mathcal{N}(0, \sigma^2 t).$$

The parameter σ is called the **volatility** of stock prices and is the basis for option pricing (Black–Scholes model). A drop of milk (millions of particles) diffusing in coffee is a macroscopic manifestation of Brownian motion.

4.4 Normal Approximation to the Binomial

We now give a detailed proof that the binomial distribution is well approximated by the normal. This is a concrete version of the central limit theorem for the special case of coin flipping. The proof uses Stirling's approximation and a ratio argument.

4.4.1 Setup

Let $X \sim \text{Binomial}(n, 1/2)$. Then $\mu = n/2$, $\sigma^2 = n/4$, $\sigma = \sqrt{n}/2$. Let

$$Z = \frac{X - \mu}{\sigma} = \frac{X - n/2}{\sqrt{n}/2}.$$

Then $\mathbb{E}(Z) = 0$, $\text{Var}(Z) = 1$, no matter what n is. Z takes discrete values with spacing $\Delta z = 1/\sigma = 2/\sqrt{n}$. We want to show:

$$p(z)/\Delta z \to f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Our strategy is in two steps. First, we show that the probability at the center ($z = 0$, i.e., $X = n/2$) matches the normal density. Then we show that the *ratio* of probabilities at z to the probability at $z = 0$ matches $e^{-z^2/2}$.

4.4.2 Step 1: The Central Term

Using Stirling's approximation $n! \sim \sqrt{2\pi n}\, n^n\, e^{-n}$:

$$\begin{aligned} P(X = n/2) &= \frac{n!}{(n/2)!^2\, 2^n} \\ &\sim \frac{\sqrt{2\pi n}\, n^n\, e^{-n}}{(\sqrt{2\pi(n/2)}\, (n/2)^{n/2})^2\, 2^n} \\ &\sim \frac{1}{\sqrt{2\pi}} \cdot \frac{2}{\sqrt{n}}. \end{aligned}$$

Let us carefully verify the cancellation in the second line. The denominator is $(\sqrt{\pi n})^2 \cdot ((n/2)^{n/2})^2 \cdot e^{-2(n/2)} \cdot 2^n = \pi n \cdot (n/2)^n \cdot e^{-n} \cdot 2^n$. Now $(n/2)^n \cdot 2^n = (n/2 \cdot 2)^n = n^n$. So the denominator becomes $\pi n \cdot n^n \cdot e^{-n}$. Dividing: $\frac{\sqrt{2\pi n} \cdot n^n \cdot e^{-n}}{\pi n \cdot n^n \cdot e^{-n}} = \frac{\sqrt{2\pi n}}{\pi n} = \frac{\sqrt{2}}{\sqrt{\pi n}} = \frac{1}{\sqrt{2\pi}} \cdot \frac{2}{\sqrt{n}}$.

Since $\Delta z = 2/\sqrt{n}$, this gives $p(0) \approx \frac{1}{\sqrt{2\pi}}\, \Delta z$, confirming the normal density at $z = 0$.

4.4.3 Step 2: The Ratio

Let $k = n/2+d$ where $d = z\sqrt{n}/2$. We compute the ratio of the binomial probability at $k = n/2+d$ to the probability at $k = n/2$:

$$\begin{aligned}\frac{P(X = n/2 + d)}{P(X = n/2)} &= \frac{\binom{n}{n/2+d}}{\binom{n}{n/2}} \\ &= \frac{(n/2)!\,(n/2)!}{(n/2+d)!\,(n/2-d)!} \\ &= \frac{(n/2)(n/2-1)\cdots(n/2-(d-1))}{(n/2+1)(n/2+2)\cdots(n/2+d)}.\end{aligned}$$

The second line uses the fact that $n!$ cancels from numerator and denominator. The third line writes the ratio of factorials as a product of fractions: the numerator has factors from $(n/2)$ down to $(n/2-d+1)$, and the denominator has factors from $(n/2+1)$ up to $(n/2+d)$.

Let $\delta = 2/n$. Each factor in the numerator is of the form $1 - j\delta$ and each in the denominator is $1 + j\delta$. Using $1 \pm j\delta \approx e^{\pm j\delta}$:

$$\begin{aligned}\frac{P(X = n/2 + d)}{P(X = n/2)} &\approx \frac{e^{-\delta}\, e^{-2\delta} \cdots e^{-(d-1)\delta}}{e^{\delta}\, e^{2\delta} \cdots e^{d\delta}} \\ &= \frac{e^{-(1+2+\cdots+(d-1))\delta}}{e^{(1+2+\cdots+d)\delta}} \\ &= \frac{e^{-d(d-1)\delta/2}}{e^{d(d+1)\delta/2}} \\ &= e^{-[d(d-1)/2+d(d+1)/2]\delta} = e^{-d^2\delta} \\ &= e^{-(z\sqrt{n}/2)^2(2/n)} = e^{-z^2/2}.\end{aligned}$$

In the third line, we use the formula $1 + 2 + \cdots + m = m(m+1)/2$. In the fourth line, $d(d-1)/2 + d(d+1)/2 = d[(d-1)+(d+1)]/2 = d \cdot 2d/2 = d^2$. Finally, $d^2\delta = (z\sqrt{n}/2)^2 \cdot (2/n) = z^2 n/(4) \cdot 2/n = z^2/2$.

Combining Steps 1 and 2: $p(z)/\Delta z \to \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$.

This completes the proof: the binomial distribution, properly normalized, converges to the standard normal density.

4.4.4 General Binomial

For $X \sim \text{Binomial}(n, p)$:

$$X \approx \mathcal{N}(np,\, np(1-p)), \qquad X/n \approx \mathcal{N}(p,\, p(1-p)/n).$$

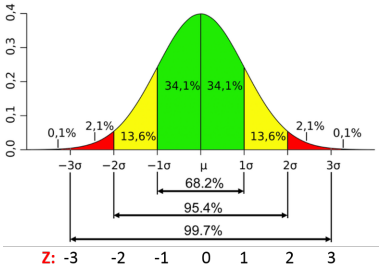


Figure 4.10: Normal approximation to the binomial.

Example ($n = 100$, $p = 1/2$): $X \sim \mathcal{N}(50, 25)$, so $P(X \in [40, 60]) = P(|X - 50| \leq 10) = P(|Z| \leq 2) \approx 95\%$. Recall $\sum_{k=40}^{60} \binom{100}{k}/2^{100} \to$ integral.

Example (polling): $n = 100$, $p = 0.2$. Then $X/n \sim \mathcal{N}(0.2, 0.04^2)$, so $P(X/n \in [0.12, 0.28]) \approx 95\%$.

Example (Monte Carlo): $n = 10000$, $p = \pi/4$. Then $4m/n \approx \mathcal{N}(\pi, \pi(4 - \pi)/10000)$.

4.5 Conditional Independence and Graphical Models

We introduced conditional independence briefly in Chapter 1. Here we develop it more systematically and connect it to graphical models, Markov decision processes, and reinforcement learning.

4.5.1 Markov Chain

$Z \to X \to Y$ (child $\perp$ grandparent given parent):

$$p(y \mid x, z) = p(y \mid x).$$

Figure 4.11: Markov chain: future is independent of past given present.

This is the Markov property we encountered in Chapter 1: once we know the present state X, the future Y is independent of the past Z. The present "screens off" the past from the future. For example, if Z = yesterday's weather, X = today's weather, and Y = tomorrow's weather, then knowing today's weather makes yesterday's weather irrelevant for predicting tomorrow. All the information from the past that is relevant to the future is captured in the present state.

4.5.2 Shared Cause

$X \leftarrow Z \to Y$ (siblings $\perp$ given parent):

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z).$$

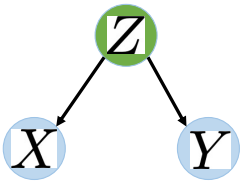


Figure 4.12: Shared cause: children are conditionally independent given parent.

The shared cause structure says: X and Y are conditionally independent given Z, because Z is the common cause of both. In the population, X and Y may appear correlated (e.g., lung cancer and bronchitis are correlated), but this correlation is entirely *explained* by the shared cause Z (smoking). Once we condition on Z, the correlation disappears.

4.5.3 Markov Decision Process

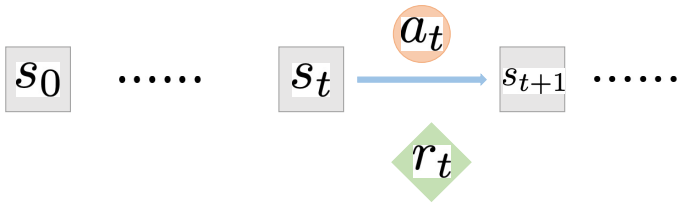


Figure 4.13: Markov decision process: state s_t, action a_t, reward r_t.

A **Markov decision process** (MDP) extends the Markov chain by adding actions and rewards. At each time step t: the system is in state s_t; the agent takes action a_t according to policy $\pi(a_t \mid s_t)$; the system transitions to state s_{t+1} via dynamics $p(s_{t+1} \mid s_t, a_t)$; the agent receives reward r_t.

The Markov property still holds: the future state depends only on the current state and action, not on the history. This makes the problem tractable — we do not need to remember the entire past, just the current state.

The **return** is the discounted sum of future rewards: $R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$ where $0 < \gamma < 1$. The discount factor γ reflects the idea that immediate rewards are worth more than distant ones. The geometric series $1 + \gamma + \gamma^2 + \cdots = 1/(1-\gamma)$ ensures the total return is finite.

The **value function** is $V(s) = \mathbb{E}[R_t \mid s_t = s]$, and the **action-value function** is $Q(s, a) = \mathbb{E}[R_t \mid s_t = s, a_t = a]$.

Reinforcement learning seeks a policy π that maximizes $V(s_0)$. Intuition: imagine 1 million people playing out the process; the value function is the average return. This connects reinforcement learning back to our fundamental theme: expectations are population averages.

4.5.4 Bayes Networks Revisited

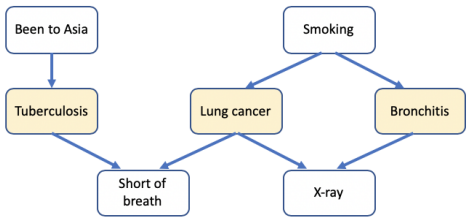


Figure 4.14: Bayes network for medical diagnosis.

$p(a, s, t, l, b, d, x) = p(a)\, p(s)\, p(t \mid a)\, p(l \mid s)\, p(b \mid s)\, p(d \mid t, l)\, p(x \mid b, l)$.

This factorization encodes the conditional independence structure of the problem. Without any independence structure, specifying the joint distribution of 7 binary variables would require $2^7 - 1 = 127$ parameters. With the Bayes network structure, we need far fewer: $p(a)$ needs 1 parameter, $p(s)$ needs 1, $p(t \mid a)$ needs 2, and so on. The total is much smaller than 127.

Inference involves summing over hidden variables, computed efficiently by message passing / belief propagation. These algorithms exploit the graph structure to avoid the exponential blowup of naively summing over all combinations.

4.6 Transformations of Random Variables

We often need to find the distribution of a transformed random variable $Y = r(X)$ when we know the distribution of X. The key principle is **conservation of probability**: the same number of people must be in corresponding intervals, regardless of how we relabel the axis.

4.6.1 Linear Change of Variable

If $X \sim f(x)$ and $Y = aX + b$ with $a > 0$:

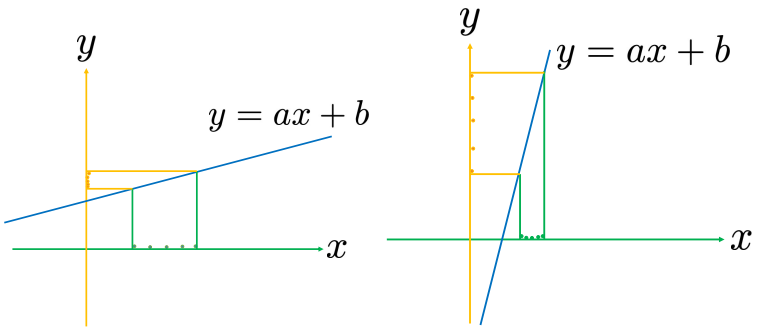


Figure 4.15: Linear transformation: stretching and shifting the density.

Conservation of probability: $P(X \in (x, x + \Delta x)) = P(Y \in (y, y + \Delta y))$, so $f(x)\,\Delta x = g(y)\,\Delta y$. Since $y = ax + b$ and $\Delta y = a\,\Delta x$:

$$g(y) = f(x)\,\frac{\Delta x}{\Delta y} = f\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}.$$

This is **space warping**: stretching or squeezing the density. When we stretch the x-axis by a factor of a (making the interval wider), the density must shrink by a factor of $1/a$ to keep the total probability equal to 1. Think of it like spreading the same amount of butter over a wider piece of toast — the layer gets thinner.

4.6.2 Nonlinear Change of Variable

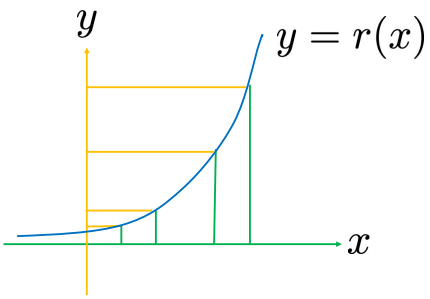


Figure 4.16: Nonlinear transformation: locally linear space warping.

For a monotone transformation $Y = r(X)$ with inverse $x = r^{-1}(y)$: $f(x)\,\Delta x = g(y)\,\Delta y$, $\Delta y/\Delta x = r'(x)$, so:

$$g(y) = f(r^{-1}(y)) \cdot \left|\frac{dx}{dy}\right| = \frac{f(r^{-1}(y))}{|r'(r^{-1}(y))|}.$$

The idea is the same as the linear case, but now the "stretching factor" varies from point to point. At each location x, the transformation is *locally* linear with slope $r'(x)$. Where the transformation stretches space ($|r'(x)| > 1$), the density gets compressed; where it compresses space ($|r'(x)| < 1$), the density gets amplified.

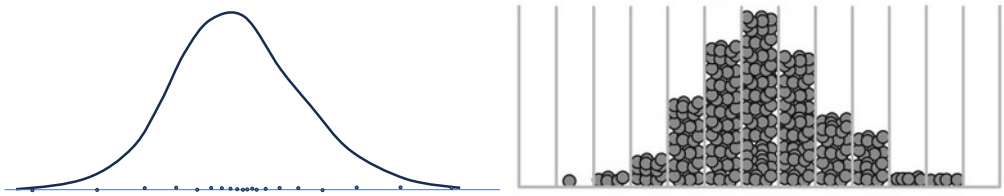


Figure 4.17: Space warping: squeezing or stretching bins changes the density.

The absolute value $|dx/dy|$ appears because density is always non-negative. Whether the transformation is increasing or decreasing, the density gets multiplied by the magnitude of the local compression/stretching factor.

4.6.3 Order-Preserving Mappings

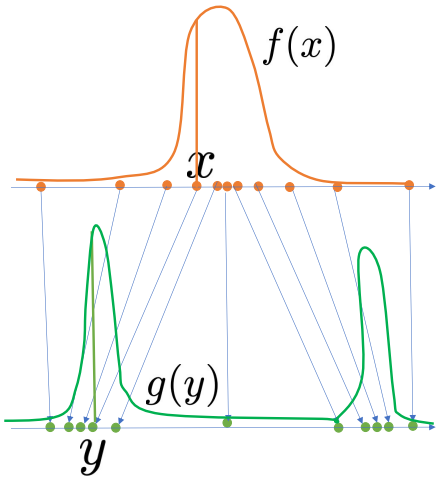


Figure 4.18: An order-preserving mapping: $F(x) = G(y)$.

If $Y = r(X)$ is monotone increasing, then $P(X \leq x) = P(Y \leq y)$, i.e., $F(x) = G(y)$.

This elegant relation says: the CDF of X evaluated at x equals the CDF of Y evaluated at the corresponding point $y = r(x)$. The percentile rank is preserved by any monotone increasing transformation. If you are at the 74th percentile of the X distribution, you are also at the 74th percentile of the Y distribution.

4.7 Simulation: Inversion and Polar Methods

How do we generate random numbers from a specific distribution on a computer? This is a fundamental problem in computational statistics, Monte Carlo simulation, and modern generative AI. The starting point is always a source of uniform random numbers, which every programming language provides.

4.7.1 The Inversion Method

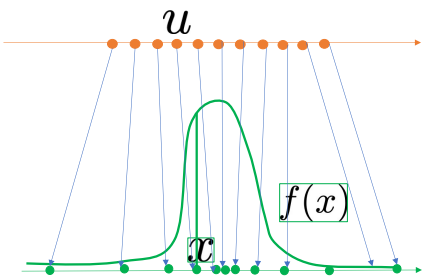


Figure 4.19: The inversion method: $U \sim \text{Uniform}[0,1]$, $X = F^{-1}(U)$.

If $U \sim \text{Uniform}[0,1]$ and F is the CDF of the target distribution, then $X = F^{-1}(U)$ has distribution F.

Intuition from the population picture: Order the population as $x_1 \le x_2 \le \cdots \le x_N$. Sample $i \sim \text{Uniform}\{1,\ldots,N\}$ and return x_i. Then $P(X \le x_i) = i/N = F(x_i)$, and $U = i/N \sim \text{Uniform}[0,1]$, so $x_i = F^{-1}(U)$.

Here is the logic in detail. Picking a random person from a population is the same as picking a random rank (a random number between 1 and N) and returning the person with that rank. The rank, divided by N, is uniform on $[0,1]$. Converting from the rank back to the value is exactly applying the inverse CDF.

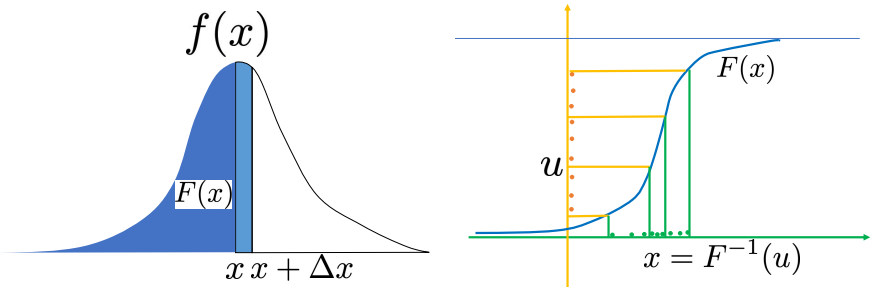


Figure 4.20: Inversion method: uniform points on the y-axis map to non-uniform points on the x-axis.

The density of $X = F^{-1}(U)$ is $f(x) = F'(x)$ because $P(U \in (u, u+\Delta u)) = P(X \in (x, x+\Delta x))$ and $\Delta u = f(x)\,\Delta x$.

Example: To generate $X \sim \text{Exponential}(1)$: $F(x) = 1 - e^{-x}$, so $x = -\log(1-u)$.

To verify: we set $u = F(x) = 1 - e^{-x}$ and solve for x. Rearranging: $e^{-x} = 1-u$, so $-x = \log(1-u)$, giving $x = -\log(1-u)$. Since $1-U$ has the same distribution as U when $U \sim \text{Uniform}[0,1]$, we can also write $x = -\log(u)$.

4.7.2 The Polar Method for Normal Random Variables

Generating normal random variables requires a clever trick because the normal CDF has no closed form. We cannot simply invert $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt$ because this integral has no elementary formula. Instead, we work in two dimensions.

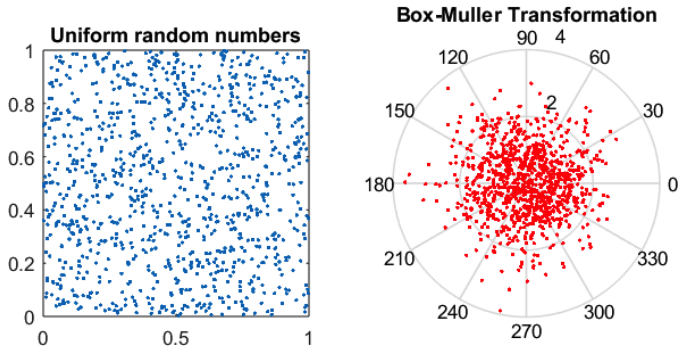


Figure 4.21: The polar method: from two uniforms to two normals.

Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ independently. Their joint density is:

$$f(x,y) = f(x)\, f(y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right).$$

Notice the beautiful structure: the joint density depends on (x,y) only through $x^2+y^2 = r^2$, which is the squared distance from the origin. This means the density is rotationally symmetric — it looks the same from every direction. This rotational symmetry is the key to the polar method.

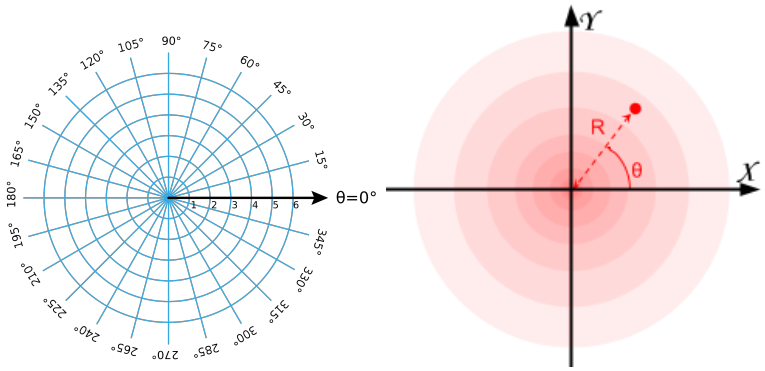


Figure 4.22: The 2D normal distribution and polar coordinates.

Switch to polar coordinates: $x = r\cos\theta$, $y = r\sin\theta$. The area of a thin ring at radius r is $2\pi r\,\Delta r$:

$$P(R \in (r, r+\Delta r)) = \frac{1}{2\pi}\, e^{-r^2/2}\, 2\pi r\,\Delta r = r\, e^{-r^2/2}\,\Delta r.$$

The 2π from the ring's circumference cancels the $1/(2\pi)$ in the density. The resulting density for R is $f_R(r) = r\, e^{-r^2/2}$.

Let $T = r^2/2$, so $\Delta T = r\,\Delta r$. Then $f(t) = e^{-t}$, meaning $T \sim$ Exponential(1). The angle $\theta \sim$ Uniform$[0, 2\pi)$, independent of R.

This is a remarkable fact: in polar coordinates, the radius (suitably transformed) and the angle are independent, and each has a simple distribution. We know how to generate exponential and uniform random variables, so we can generate normal random variables!

Algorithm: Generate $U_1, U_2 \sim$ Uniform$[0,1]$ independently. Set $T = -\log(1-U_1)$, $R = \sqrt{2T}$, $\theta = 2\pi U_2$. Then $X = R\cos\theta$ and $Y = R\sin\theta$ are independent standard normals. This transforms $(U_1, U_2) \to (X, Y)$.

This is known as the Box–Muller transform. It converts two uniform random variables into two independent standard normal random variables — a beautiful application of the change-of-variable technique.

4.7.3 Generative Models via Nonlinear Transformation

Figure 4.23: A generative model: transform Gaussian noise into realistic images via a learned neural network.

Modern generative models extend this idea: start with X consisting of iid Gaussian noise, apply a nonlinear transformation $Y = r(X)$ learned by a neural network. The function r warps the simple Gaussian distribution into a complex distribution over images, audio, or text.

The polar method shows us that even the normal distribution can be generated by transforming simpler distributions. Generative AI takes this principle to its logical extreme: learn a transformation that maps simple noise into complex, realistic data. The same mathematical framework — probability distributions and their transformations — underlies both.

4.8 Convexity and the Jensen Inequality

Jensen's inequality is one of the most useful tools in probability, with applications ranging from information theory to economics to machine learning. It answers a simple question: how does the expectation of a function compare to the function of the expectation?

4.8.1 Expectation of a Function vs. Function of Expectation

A fundamental question: how does $\mathbb{E}[h(X)]$ compare to $h(\mathbb{E}(X))$?

For a **linear** function $h(x) = ax + b$: $\mathbb{E}[h(X)] = a\mathbb{E}(X) + b = h(\mathbb{E}(X))$. They are equal. This is just the linearity of expectation.

For a **nonlinear** function, they generally differ. For $h(x) = x^2$:

$$\mathbb{E}[h(X)] - h(\mathbb{E}(X)) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \mathrm{Var}(X) \geq 0.$$

So $\mathbb{E}[h(X)] \geq h(\mathbb{E}(X))$ when h is convex.

This example with $h(x) = x^2$ is not a coincidence — it is a special case of a general principle. The inequality $\mathbb{E}[h(X)] \geq h(\mathbb{E}(X))$ holds for *any* convex function h, and the gap is related to how much X varies (the variance).

4.8.2 Convex Functions and Supporting Lines

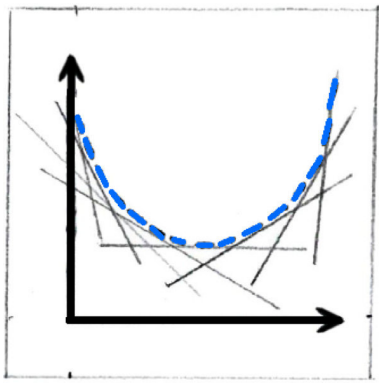


Figure 4.24: A convex function and its supporting lines.

A function h is **convex** if the line segment between any two points on its graph lies above (or on) the graph. Equivalently, h has a **supporting line** at every point x_0: a line $ax + b$ that touches h at x_0 and lies below h everywhere else.

Familiar convex functions include x^2, e^x, $|x|$, and $-\log x$ (for $x > 0$). A function is convex if and only if $h''(x) \geq 0$ everywhere (assuming h is twice differentiable). The supporting line at x_0 is just the tangent line: $h(x_0) + h'(x_0)(x - x_0)$.

4.8.3 The Jensen Inequality

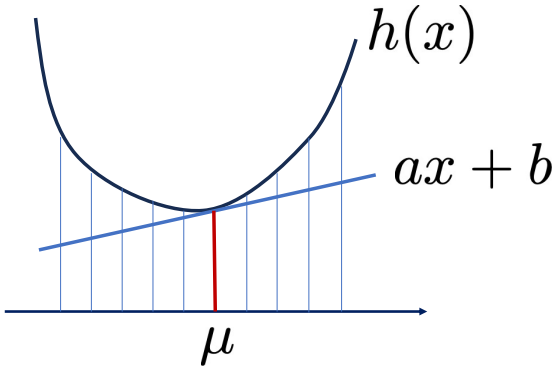


Figure 4.25: Jensen's inequality: $\mathbb{E}[h(X)] \geq h(\mathbb{E}(X))$ for convex h.

Theorem 4.8.1 (Jensen's inequality)**.** *If h is convex, then $\mathbb{E}[h(X)] \geq h(\mathbb{E}(X))$.*

Proof. Let $\mu = \mathbb{E}(X)$. At the point μ, the supporting line is $h(\mu) = a\mu + b$ and $h(x) \geq ax + b$ for all x. Therefore:

$$\begin{aligned} \mathbb{E}[h(X)] &\geq \mathbb{E}[aX + b] \qquad \text{(since } h(X) \geq aX + b \text{ for every } X\text{)} \\ &= a\,\mathbb{E}(X) + b \qquad \text{(linearity of } \mathbb{E}\text{)} \\ &= a\mu + b = h(\mu) = h(\mathbb{E}(X)). \end{aligned}$$

□

The proof is elegant: the supporting line provides a linear lower bound on h, and linear functions commute with expectation. Since $h(X) \geq aX + b$ pointwise, taking expectations preserves the inequality, and the right-hand side simplifies to $h(\mu)$.

The inequality also has a data interpretation. For data points $x_1, \ldots, x_n$:

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \geq h\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = h(\bar{x}).$$

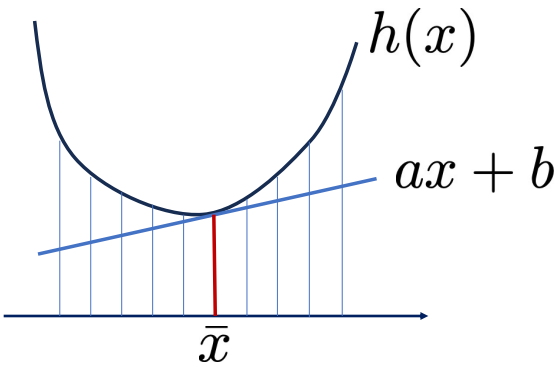


Figure 4.26: Jensen's inequality for data: the average of $h(x_i)$ exceeds h of the average.

This data version is easy to visualize: plot the data points on the graph of h. The average of their y-coordinates (the left side) lies on or above the y-coordinate of their average x-coordinate (the right side), because the curve bows upward.

4.8.4 Application to Utility and Risk

Consider two offers with the same expected dollar value $\mu = \mathbb{E}(X)$:

Offer 1: Get $\$\mu$ with certainty (utility $= h(\mu)$).

Offer 2: Get $\$X \sim f(x)$ with $\mathbb{E}(X) = \mu$ (expected utility $= \mathbb{E}[h(X)]$).

If h is **convex**: $\mathbb{E}[h(X)] \geq h(\mu)$, so you prefer Offer 2 — you are **risk-taking**. The gamble has higher expected utility than the sure thing.

If h is **concave**: $\mathbb{E}[h(X)] \leq h(\mu)$, so you prefer Offer 1 — you are **risk-averse**. You would rather have the guaranteed amount than gamble, even when the expected dollar amounts are the same.

For $h(x) = -x^2$ (concave): the gap is $h(\mu) - \mathbb{E}[h(X)] = \mathbb{E}(X^2) - \mu^2 = \mathrm{Var}(X) \geq 0$.

Most people are risk-averse for large stakes (the utility of going from \$0 to \$1 million is much larger than going from \$1 million to \$2 million), which is why insurance companies exist. The concavity of the utility function quantifies this risk aversion.

4.9 Entropy and Information Theory

Our final topic connects probability to the theory of information. How much "information" does a random variable carry? How efficiently can we encode it? These questions, first answered by Claude Shannon in 1948, led to the field of information theory, which underlies all modern communication and data compression.

4.9.1 Entropy as Expected Code Length

Consider a random variable X with four possible values:

x	A	B	C	D
$p(x)$	1/2	1/4	1/8	1/8
$-\log_2 p(x)$	1	2	3	3
coin code	H	TH	TTH	TTT

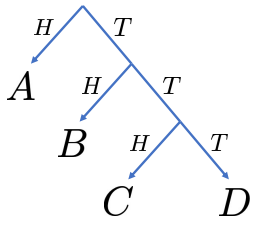


Figure 4.27: A binary tree encoding: more probable symbols get shorter codes.

We encode each symbol using coin flips: A $\to$ H (1 flip), B $\to$ TH (2 flips), C $\to$ TTH (3 flips), D $\to$ TTT (3 flips). The code length of symbol x is $l(x) = -\log_2 p(x)$. This is a **prefix code**: no codeword is a prefix of another, so decoding is unambiguous.

Why is the code length $-\log_2 p(x)$? Because a symbol with probability p carries $-\log_2 p$ bits of information. A very likely event (like the sun rising tomorrow) is not very surprising and carries little information. A very unlikely event (like winning the lottery) is highly surprising and carries a lot of information. The $-\log_2$ function captures this: smaller probabilities give larger code lengths.

Definition 4.9.1 (Entropy)**.** The **entropy** of a distribution p is the expected code length:

$$\mathbb{H}(p) = \mathbb{E}_p[-\log_2 p(X)] = \sum_x (-\log_2 p(x))\, p(x).$$

For our example: $\mathbb{H}(p) = 1 \times 1/2 + 2 \times 1/4 + 3 \times 1/8 + 3 \times 1/8 = 1.75$ flips.

Entropy measures **average surprise** or **uncertainty**. A distribution concentrated on a single outcome has entropy 0 (no surprise — we always know what will happen). A uniform distribution has maximum entropy (maximum uncertainty — every outcome is equally surprising). The shortest possible code produces completely random sequences that cannot be compressed further.

4.9.2 Kullback–Leibler Divergence

What happens if we use the wrong distribution q to design our code? If we code using q but the true distribution is p, the expected code length is:

$$\mathbb{E}_p[-\log_2 q(X)] = \sum_x (-\log_2 q(x))\, p(x).$$

This is longer than the optimal $\mathbb{H}(p)$. The **redundancy** is the **KL divergence**:

$$\mathbb{D}_{\mathrm{KL}}(p\|q) = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \sum_x \left(\log \frac{p(x)}{q(x)}\right) p(x) \geq 0.$$

The KL divergence measures how much extra code length we incur by using q instead of p. It equals zero if and only if $p = q$.

The KL divergence is not a true "distance" in the mathematical sense because it is not symmetric: $\mathbb{D}_{\mathrm{KL}}(p\|q) \neq \mathbb{D}_{\mathrm{KL}}(q\|p)$ in general. Nevertheless, it is one of the most important quantities in statistics and machine learning.

4.9.3 Non-Negativity via Jensen's Inequality

We prove $\mathbb{D}_{\mathrm{KL}}(p\|q) \geq 0$ using Jensen's inequality. First observe:

$$\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] = \sum_x \frac{q(x)}{p(x)}\, p(x) = \sum_x q(x) = 1.$$

This step is simple but worth pausing on: the $p(x)$ in the denominator cancels with the $p(x)$ from the expectation, leaving just $\sum_x q(x) = 1$.

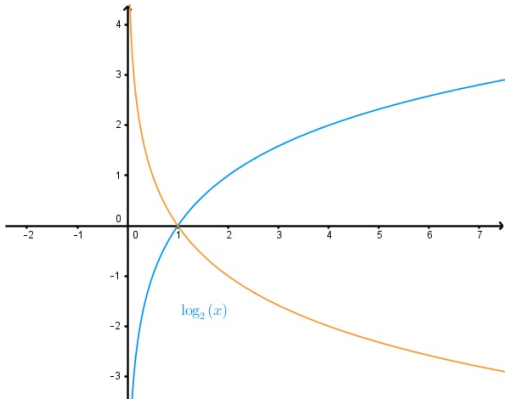


Figure 4.28: The function $-\log$ is convex.

Since $-\log$ is a convex function, Jensen's inequality gives:

$$\begin{aligned}
\mathbb{D}_{\mathrm{KL}}(p\|q) &= \mathbb{E}_p\left[-\log\frac{q(X)}{p(X)}\right] \\
&\geq -\log \mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] \qquad \text{(Jensen: } \mathbb{E}[h(X)] \geq h(\mathbb{E}(X)) \text{ for convex } h\text{)} \\
&= -\log 1 = 0.
\end{aligned}$$

Equality holds if and only if $q(X)/p(X)$ is constant, which requires $p = q$.

This proof is a beautiful application of Jensen's inequality. The convexity of $-\log$ gives us the non-negativity of KL divergence "for free."

The KL divergence is fundamental in machine learning: training a model q_θ to approximate a true distribution p is equivalent to minimizing $\mathbb{D}_{\mathrm{KL}}(p\|q_\theta)$, which is equivalent to maximizing the log-likelihood $\mathbb{E}_p[\log q_\theta(X)]$. This connects the abstract theory of information to the practical task of fitting models to data, and provides the theoretical justification for maximum likelihood estimation — one of the most widely used methods in statistics.