Probability

Lecture Notes for Stat-134, UC Berkeley
(based on R. Durret and J. Pitman)

Drafted for internal teaching use by Ziran Liu

06/23/2024

These notes are written as an organized material only for teaching STAT-134, not for any other purpose (don't spread). They are not official course materials. Their scope is intentionally aligned with the usual Stat 134 themes: probability axioms, conditional probability, independence, discrete and continuous random variables, conditional expectation, laws of large numbers, the central limit theorem, and selected topics such as characteristic functions, the Poisson process, and Markov chains. The writing aims to be rigorous, detailed, and friendly to strong undergraduates seeing the subject seriously for the first time.

Preface

Probability is one of the central languages of modern quantitative thought. It is the grammar behind statistical inference, stochastic modeling, machine learning, signal processing, queueing systems, finance, genetics, and much of theoretical computer science. Yet probability is also a surprisingly delicate subject pedagogically. Students often meet it first through quick computations, recipes, and mnemonic formulas. Those tools are useful, but they can obscure the deeper architecture of the subject: probability begins with a model, proceeds by mathematical structure, and only then yields formulas.

These notes were written to support a full-semester undergraduate course in the spirit of UC Berkeley's Stat 134: a course that is conceptually serious, computation-rich, and broad enough to prepare students for later work in statistics, probability, stochastic processes, and mathematical data science. The presentation is intentionally detailed. Definitions are motivated before they are formalized. Theorems are followed by interpretation. Proofs are included whenever they materially improve understanding, and when a proof would take us too far afield, the notes say so explicitly and explain what the omitted argument would need.

The audience I have in mind is a student with one year of calculus and a willingness to reason carefully. Measure-theoretic language appears when it clarifies the subject, but the main line of the text stays anchored in calculational and conceptual probability. The result is neither a purely informal first course nor a fully abstract graduate treatment. Instead, it is an ambitious bridge: rigorous enough to withstand scrutiny, but paced so that a diligent undergraduate can genuinely learn from it.

The manuscript is organized in four layers. The first develops foundations: sample spaces, axioms, counting, conditional probability, independence, and the standard discrete models. The second develops expectation, variance, continuous models, joint distributions, transformations, and conditional expectation. The third develops asymptotics through laws of large numbers, normal approximation, generating functions, and characteristic functions. The fourth turns to selected stochastic processes, emphasizing the Poisson process and Markov chains, along with a short optional chapter on further topics. Each chapter ends with exercises, and cumulative review sets are inserted periodically to support exam preparation.

Two books influenced the architecture of these notes in different ways. Jacod and Protter provide a clean path from probability axioms to characteristic functions, convergence, conditional expectation, and martingale ideas. Durrett provides a broader probability landscape, especially for

limit theorems, Poisson processes, and Markov chains. I have not attempted to imitate either text. Instead, I have tried to produce a new undergraduate narrative that keeps the conceptual precision of the first and the breadth of the second while remaining teachable in a single semester.

A final comment for instructors: this is a deliberately overprepared manuscript. In a live course, not every theorem here needs equal emphasis. Some sections are intended as core lecture material; others are best treated in discussion, in homework, or as reading. A suggested pacing guide appears after the table of contents. If you teach from these notes, you should feel free to omit a few advanced proofs while keeping the main conceptual arc intact.

Contents

Suggested Course Pacing

A Berkeley-style Stat 134 course is typically organized around three lecture hours and two discussion hours per week. The following pacing guide assumes roughly fourteen instructional weeks plus final-review time. It is a suggestion, not a prescription.

Week	Core chapters	Suggested emphasis
1	Chapters 1–2	Modeling, axioms, counting, conditional probability, Bayes' rule.
2	Chapters 2–3	Independence, Bernoulli trials, discrete random variables, cdfs.
3	Chapters 3–4	Standard discrete families, expectation, indicator methods.
4	Chapter 4	Variance, covariance, inequalities, first cumulative review.
5	Chapter 5	Continuous random variables, densities, exponential and normal models.
6	Chapters 6–7	Joint distributions, conditioning, transformations, order statistics.
7	Chapter 8	Conditional expectation, tower property, prediction viewpoint.
8	Chapter 9	Generating functions, moment generating functions, characteristic functions.
9	Chapter 10	Modes of convergence, weak and strong laws of large numbers.
10	Chapter 11	Central limit theorem, normal approximation, continuity correction.
11	Chapters 12–13	Poisson approximation and the Poisson process.
12	Chapter 14	Markov chains, transition matrices, first-step analysis.
13	Chapter 14	Stationary distributions, long-run behavior, gambler's ruin.
14	Chapter 15	Optional further topics, cumulative review, exam preparation.

Chapter 1

Probability Models, Events, and Counting

Probability is not just a collection of formulas. It is a way of building mathematical models for uncertain situations. A good probability calculation begins by answering three questions: what are the possible outcomes, which collections of outcomes count as events, and what probability should each event receive?

1.1 Why probability begins with modeling

Probability theory studies uncertainty through mathematical models. This sounds obvious, but it is easy to underestimate how important the word *model* is. The world itself does not come equipped with a sample space or a list of events. We create those objects. When we toss a coin twice, do we care about the exact angular momentum of the coin in the air? Usually not. We simplify the world to the four outcomes HH, HT, TH, TT. When we observe a patient's test result, do we model the patient's entire medical history, or only whether the patient has a disease and whether the test is positive? Different questions require different models.

The quality of a probability answer depends on the quality of the model underneath it. A perfectly correct calculation inside a poor model can still lead to a poor scientific conclusion. For this reason, the first habit of mind we want to develop is not computation but representation. Before calculating, pause and ask: what is the experiment? What are the outcomes? Which details matter and which do not? Which symmetries justify equal probabilities, and which do not?

A probability model has three basic pieces.

Definition 1.1. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

(i) Ω is the *sample space*, the set of all outcomes under consideration;

(ii) $\mathcal{F}$ is a collection of subsets of Ω, called *events*;

(iii) $\mathbb{P}$ is a function assigning a number in $[0, 1]$ to each event in $\mathcal{F}$.

In a first course, most of our sample spaces will be finite, countable, or subsets of $\mathbb{R}$. In such settings, one can often think of $\mathcal{F}$ simply as "the events we are allowed to talk about." Later, especially in continuous models, $\mathcal{F}$ must be chosen more carefully. For now, the main point is practical: probability is assigned to events, not directly to individual numerical values, and events are sets.

Example 1.2 (Two coin tosses)**.** Suppose we toss a fair coin twice and record the ordered sequence of results. Then

$$\Omega = \{HH, HT, TH, TT\}.$$

An event is any subset of Ω. For instance,

$$A = \{HH, HT\} \quad \text{(first toss is heads)}, \qquad B = \{HT, TH\} \quad \text{(exactly one head)}.$$

If the coin is fair and the two tosses are modeled as symmetric, then each outcome has probability $1/4$.

The ordered sequence matters here. If instead we cared only about the number of heads, then the outcome space could be $\{0, 1, 2\}$. Neither choice is "more correct" in the abstract; one is more useful for a given question.

1.1.1 Events are sets, and set language matters

The set-theoretic operations on events correspond directly to logical operations in probability.

- $A \cup B$ means "A or B occurs" (at least one of them occurs).
- $A \cap B$ means "A and B occur."
- A^c means "A does not occur."
- $A \setminus B$ means "A occurs but B does not."

This language is not decoration. It is the grammar through which nearly all probability identities are written.

A common beginner's mistake is to confuse an event with a numerical summary of an event. For instance, in a dice experiment the event "the roll is even" is a subset $\{2, 4, 6\}$ of the sample space $\{1, 2, 3, 4, 5, 6\}$; it is not the number 2 just because there are three even outcomes. Keeping this distinction clear avoids many later confusions.

1.2 The axioms of probability

The modern foundation of probability is based on a small number of axioms. They are deliberately sparse. The idea is that a few natural rules should imply a large and useful theory.

Definition 1.3. A function $\mathbb{P}$ on events is a *probability measure* if it satisfies:

(P1) $\mathbb{P}(A) \geq 0$ for every event A;

(P2) $\mathbb{P}(\Omega) = 1$;

(P3) for any countable collection of pairwise disjoint events $A_1, A_2, \ldots$,

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

The third property is called *countable additivity*.

Countable additivity deserves attention. If events cannot occur together, then the probability that one of them occurs should equal the sum of their probabilities. For a finite list of disjoint events this sounds obvious. The countable version is a deeper principle and one of the great strengths of the axiomatic framework. It is what makes infinite processes mathematically manageable.

A number of useful facts follow immediately.

Proposition 1.4 (Basic identities)**.** *For any events A and B,*

(a) $\mathbb{P}(\varnothing) = 0$*;*

(b) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*;*

(c) if $A \subseteq B$*, then* $\mathbb{P}(A) \leq \mathbb{P}(B)$*;*

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

Proof. For (a), note that Ω and $\varnothing$ are disjoint and $\Omega \cup \varnothing = \Omega$. Thus

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \mathbb{P}(\varnothing),$$

so $\mathbb{P}(\varnothing) = 0$.

For (b), the events A and A^c are disjoint and their union is Ω. Hence

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

For (c), if $A \subseteq B$ then $B = A \cup (B \setminus A)$ with the two pieces disjoint. Therefore

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A).$$

For (d), write $A \cup B$ as the disjoint union

$$A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B).$$

Adding probabilities of the disjoint pieces and regrouping yields the formula. □

The identity in part (d) is the two-set version of *inclusion–exclusion*. It is one of the most frequently used formulas in elementary probability. The subtraction term corrects the double counting of the overlap $A \cap B$.

Corollary 1.5 (Union bound)**.** *For any events $A_1, \dots, A_n$,*

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) \le \sum_{i=1}^{n} \mathbb{P}(A_i).$$

More generally, the same inequality holds for countably many events.

Proof. Replace the events by disjoint pieces:

$$A_1, \quad A_2 \setminus A_1, \quad A_3 \setminus (A_1 \cup A_2), \quad \dots$$

The probability of each piece is at most the probability of the corresponding original event. Summing gives the result. □

The union bound is sometimes called the *first moment method for events*. It is crude but surprisingly effective. If several bad things could happen, the probability that at least one happens is at most the sum of the separate probabilities. This upper bound is often enough for asymptotic arguments and algorithmic estimates.

Example 1.6 (At least one birthday match)**.** Suppose A_{ij} is the event that people i and j in a room of n people share a birthday, under the simplifying assumption of independent uniformly distributed birthdays among 365 days. Then

$$\mathbb{P}(A_{ij}) = \frac{1}{365}.$$

There are $\binom{n}{2}$ such pairs, so

$$\mathbb{P}(\text{some shared birthday}) \le \binom{n}{2}\frac{1}{365}.$$

This bound is not exact, but it already predicts that birthday matches become likely once $\binom{n}{2}$ is comparable to 365.

1.3 Finite sample spaces and uniform models

Many early examples in probability are based on a finite sample space with equally likely outcomes. In such models, probability reduces to counting.

Definition 1.7. If Ω is finite and each outcome $\omega \in \Omega$ has the same probability, then the model is called *uniform*. In that case,

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

for every event $A \subseteq \Omega$.

Uniform models are attractive because they convert probability questions into counting questions. But the word *uniform* must be justified, not assumed. A common trap is to declare outcomes equally likely when the experiment does not support that claim.

Example 1.8 (Cards)**.** A card is drawn uniformly from a standard deck of 52 cards. If A is the event "the card is a heart" and B is the event "the card is a face card," then

$$\mathbb{P}(A) = \frac{13}{52} = \frac{1}{4}, \qquad \mathbb{P}(B) = \frac{12}{52} = \frac{3}{13}.$$

The event $A \cap B$ consists of the jack, queen, and king of hearts, so

$$\mathbb{P}(A \cap B) = \frac{3}{52}.$$

Later we will use this same example to discuss independence.

The central combinatorial tools are permutations and combinations.

Definition 1.9. For a positive integer n, the factorial of n is

$$n! = n(n-1)(n-2)\cdots 2 \cdot 1.$$

The number of ordered arrangements of n distinct objects is $n!$.

Proposition 1.10 (Counting permutations and selections)**.** *(a) The number of ordered k-tuples of distinct elements chosen from n distinct objects is*

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

(b) The number of unordered k-element subsets chosen from n distinct objects is

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

Proof. For (a), there are n choices for the first position, $n-1$ for the second, and so on, giving

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

For (b), each unordered k-subset can be ordered in $k!$ ways, so divide the count from part (a) by $k!$. □

It is useful to interpret $\binom{n}{k}$ as "the number of ways to choose k objects from n without regard to order." This is not merely a mnemonic. It tells you when combinations, rather than permutations, are the correct tool.

1.3.1 Sampling with and without replacement

Many mistakes in counting come from failure to distinguish these two settings.

- **With replacement:** after each draw, the chosen object is returned. Different draws are often modeled as independent.
- **Without replacement:** the available pool changes after each draw. Independence usually fails.

As a simple example, suppose two balls are drawn from an urn containing 3 red and 2 blue balls.

- If we sample *with replacement*, the probability of red on each draw is always 3/5.
- If we sample *without replacement*, the probability of red on the second draw depends on the first draw.

This distinction will later reappear in the difference between binomial and hypergeometric distributions.

1.4 Multistage experiments and tree thinking

When an experiment unfolds in stages, it is often useful to represent the possibilities using a tree. The tree diagram itself is not the mathematics, but it is a highly effective organizational device. Each path through the tree corresponds to a complete outcome. The probability of a path is found by multiplying the probabilities along the branches, provided those branch probabilities are interpreted conditionally on the previous stages.

Example 1.11 (A simple two-stage experiment)**.** A box contains 2 gold coins and 3 silver coins. We draw one coin, record its type, replace it, and then draw again. The probability of two gold coins is

$$\frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25}.$$

The probability of exactly one gold coin is

$$\frac{2}{5} \cdot \frac{3}{5} + \frac{3}{5} \cdot \frac{2}{5} = \frac{12}{25}.$$

Even before we formally define conditional probability in the next chapter, the tree picture encourages the right multiplication-and-addition structure.

Tree thinking also teaches an important general lesson: probability calculations often decompose into two moves.

(1) Break the event of interest into simpler disjoint cases.

(2) Sum the probabilities of those cases.

This add-over-cases principle will later become the law of total probability.

1.5 How to choose a sample space well

A sample space should be rich enough to answer the question but no richer than necessary. This is partly a matter of taste, but mostly a matter of mathematical efficiency.

Here are some practical guidelines.

(1) **Respect order when order matters.** If three cards are drawn in sequence and the order matters, outcomes should record an ordered triple.

(2) **Avoid hidden asymmetry.** If you want to use a uniform model, make sure the chosen outcomes are truly symmetric under the physical mechanism.

(3) **Choose outcomes before choosing events.** It is tempting to define outcomes so that the target event looks simple, but this can distort the rest of the model.

(4) **Distinguish observations from latent mechanisms.** In some problems we model only what is observed; in others we model a hidden process as well. Both are legitimate, but they answer different kinds of questions.

Remark 1.12 (A famous warning)**.** Suppose a family has two children and we are told that at least one child is a girl. What is the probability both children are girls? The naive answer $1/2$ is incorrect in the standard model. If we take

$$\Omega = \{BB, BG, GB, GG\}$$

with all outcomes equally likely, then the condition "at least one girl" leaves three possibilities: BG, GB, GG. Only one of them has two girls, so the answer is $1/3$. The mistake in the naive answer comes from forgetting to specify the sample space precisely.

1.6 Summary

This chapter introduced the foundational vocabulary of probability. The central messages are worth repeating.

- Probability is a mathematical model for uncertainty, not a bag of disconnected formulas.
- Events are sets. Their algebra mirrors logical reasoning.
- The axioms of probability are simple but powerful; many working identities flow from them.
- On a finite uniform sample space, probability reduces to counting.
- Good modeling decisions are as important as correct calculations.

The next chapter builds on this foundation by introducing conditional probability, Bayes' rule, and independence. Those ideas are the true engine of probabilistic reasoning.

Exercises

Exercise 1.1. A fair die is rolled twice. Write down a sample space that records the ordered pair of outcomes. Let A be the event that the sum is at least 10, and let B be the event that the first roll is larger than the second. Compute $\mathbb{P}(A)$, $\mathbb{P}(B)$, and $\mathbb{P}(A \cap B)$.

Exercise 1.2. A card is drawn uniformly from a standard deck. Let A be the event "the card is red" and B the event "the card is a queen or king." Compute $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(A \cap B)$, and $\mathbb{P}(A \cup B)$.

Exercise 1.3. Prove the identity

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

for all events A and B.

Exercise 1.4. Suppose $A \subseteq B \subseteq C$. Show that

$$\mathbb{P}(C \setminus A) = \mathbb{P}(C \setminus B) + \mathbb{P}(B \setminus A).$$

Interpret the identity in words.

Exercise 1.5. A committee of 4 is chosen uniformly from 7 women and 5 men. What is the probability that the committee contains exactly 2 women? At least 3 women?

Exercise 1.6. How many distinct rearrangements are there of the letters in the word `STATISTICS`? Explain carefully why your counting formula is correct.

Exercise 1.7. An urn contains 6 balls numbered 1 through 6. Three balls are drawn without replacement. Give two different reasonable sample spaces for this experiment: one in which order matters and one in which order does not. Under each choice, describe what would count as a uniform model.

Exercise 1.8. Let $A_1, \dots, A_n$ be events. Prove by induction that

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) \le \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Then explain why the argument extends to a countable collection of events.

Challenge Exercise 1.9. Three points are chosen independently and uniformly on a circle. What is the probability that all three points lie on some semicircle? Start by choosing a convenient way to describe the event.

Challenge Exercise 1.10. Suppose you are told that a random positive integer is chosen "uniformly." Explain why this statement is mathematically problematic. What property of probability measures on countable sets makes a truly uniform distribution on the positive integers impossible?

Chapter 2

Conditional Probability, Bayes' Rule, and Independence

Conditional probability is probability viewed through information. When we learn that some event B has occurred, the relevant universe is no longer the full sample space but the smaller world B. Independence is the special situation in which this extra information does not change probabilities.

2.1 Conditioning as renormalization

Suppose we perform an experiment and then learn that an event B has occurred. If we now ask for the chance of another event A, the original probability $\mathbb{P}(A)$ is no longer the right quantity. Outcomes outside B are now impossible, at least relative to our information. We should therefore restrict attention to B and rescale so that the total probability of B becomes 1.

This idea leads directly to the definition.

Definition 2.1. If $\mathbb{P}(B) > 0$, the *conditional probability* of A given B is

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

The numerator counts the probability that both A and B occur. The denominator renormalizes by the size of the world in which we now know we are living. This is not merely a formula to memorize. It is a principle: condition by restricting to the available information and then rescaling.

Example 2.2 (A card example)**.** A card is drawn uniformly from a standard deck. Let A be the event "the card is a queen" and B the event "the card is a face card." Then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{4/52}{12/52} = \frac{1}{3}.$$

Once we know the card is a face card, the relevant possibilities are the 12 face cards, of which 4 are queens.

Example 2.3 (A draw without replacement)**.** An urn contains 5 red balls and 3 blue balls. Two balls are drawn without replacement. Let A be the event "the second ball is red" and B the event "the first ball is blue." Then

$$\mathbb{P}(A \mid B) = \frac{5}{7}.$$

Indeed, if the first ball is blue, the urn now contains 5 red and 2 blue balls.

Notice how naturally the answer in the second example comes from the story, even before writing down the formula. This is a good sign. The formula should explain the reasoning, not replace it.

2.1.1 Multiplication rule

The definition of conditional probability can be rearranged to yield one of the basic workhorses of the subject.

Proposition 2.4 (Multiplication rule)**.** *If $\mathbb{P}(B) > 0$, then*

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

Similarly, if $\mathbb{P}(A) > 0$,

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

This says: *probability of both = probability of the first × probability of the second given the first.* For multistage experiments, this becomes a chain rule.

Corollary 2.5 (Chain rule)**.** *If $A_1, \dots, A_n$ are events with the relevant conditional probabilities defined, then*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i\Big) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2)\cdots\mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

The chain rule is especially useful when an experiment unfolds sequentially. It formalizes what tree diagrams suggest intuitively.

Example 2.6 (Three aces in a row)**.** Three cards are drawn without replacement from a shuffled deck. The probability that all three are aces is

$$\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50}.$$

Each factor is a conditional probability given what happened before.

2.2 Partitions and the law of total probability

Often an event can occur in several mutually exclusive ways. This leads to the next central idea.

Definition 2.7. A collection of events $B_1, B_2, \dots$ is a *partition* of the sample space if the events are pairwise disjoint and

$$\bigcup_i B_i = \Omega.$$

A partition represents a complete list of possible cases. Once we condition on which case occurred, many problems become much simpler.

Theorem 2.8 (Law of total probability)**.** *Let $B_1, B_2, \dots$ be a finite or countable partition of Ω with $\mathbb{P}(B_i) > 0$ for each i. Then for any event A,*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Proof. Since the B_i form a partition,

$$A = \bigcup_i (A \cap B_i)$$

as a disjoint union. Therefore,

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

□

The theorem says that if we can split the world into manageable cases, then we can compute by averaging conditional probabilities over those cases.

Example 2.9 (A mixture of coins)**.** A box contains two coins. One is fair, the other lands heads with probability $3/4$. A coin is selected uniformly from the box and tossed once. What is the probability of heads?

Let B_1 be the event that the fair coin is chosen and B_2 the event that the biased coin is chosen. Then

$$\mathbb{P}(H) = \mathbb{P}(H \mid B_1)\mathbb{P}(B_1) + \mathbb{P}(H \mid B_2)\mathbb{P}(B_2) = \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} = \frac{5}{8}.$$

This example already reveals the logic of *mixture models*. A distribution can arise from several hidden subpopulations, and the overall probability is the weighted average across them.

2.3 Bayes' rule

The law of total probability is often used together with the multiplication rule to reverse conditions. This is Bayes' rule.

Theorem 2.10 (Bayes' rule)**.** *Let $B_1, \dots, B_n$ be a partition of Ω with $\mathbb{P}(B_i) > 0$. If A is an event with $\mathbb{P}(A) > 0$, then*

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Proof. By the multiplication rule,

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)}.$$

Now use the law of total probability to write

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

□

Bayes' rule is the mathematical form of a very common kind of reasoning: start with prior beliefs about possible causes, then update those beliefs after observing evidence. The denominator is often called the *evidence* or *normalizing constant*. The numerator is "likelihood times prior."

Example 2.11 (Medical testing)**.** Suppose 1% of a population has a disease. A test has sensitivity 0.98, meaning it returns positive with probability 0.98 on diseased individuals, and specificity 0.95, meaning it returns negative with probability 0.95 on healthy individuals. If a randomly chosen person tests positive, what is the probability the person actually has the disease?

Let D be the event "person has disease" and $+$ the event "test is positive." Then

$$\mathbb{P}(D) = 0.01, \qquad \mathbb{P}(+ \mid D) = 0.98, \qquad \mathbb{P}(+ \mid D^c) = 0.05.$$

Bayes' rule gives

$$\mathbb{P}(D \mid +) = \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.165.$$

So even with a reasonably accurate test, a positive result need not imply high posterior probability when the condition is rare.

This example is worth dwelling on. The posterior probability is not determined by the test quality alone. The base rate, $\mathbb{P}(D)$, matters enormously. Many public misunderstandings of risk are failures to account for base rates.

2.4 Independence of events

We now turn to one of the most important and most frequently misunderstood concepts in probability.

Definition 2.12. Two events A and B are *independent* if

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

A collection of events $A_1, \dots, A_n$ is *mutually independent* if for every nonempty subset $I \subseteq \{1, \dots, n\}$,

$$\mathbb{P}\Big(\bigcap_{i \in I} A_i\Big) = \prod_{i \in I} \mathbb{P}(A_i).$$

The equation for two events says that the probability of both events occurring factors as the product of their separate probabilities. When $\mathbb{P}(B) > 0$, this is equivalent to

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

So independence means exactly that learning B gives no information about whether A occurs.

Remark 2.13. Independence is not the same as disjointness. In fact, if A and B are disjoint and both have positive probability, then they cannot be independent, because $\mathbb{P}(A \cap B) = 0$ while $\mathbb{P}(A)\mathbb{P}(B) > 0$.

Example 2.14 (A standard independent model)**.** Toss a fair coin twice. Let A be the event "first toss is heads" and B the event "second toss is heads." Then

$$\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}, \qquad \mathbb{P}(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2},$$

so the events are independent.

Example 2.15 (Not independent)**.** Draw two cards without replacement from a deck. Let A be the event "first card is an ace" and B the event "second card is an ace." Then

$$\mathbb{P}(A) = \frac{4}{52}, \qquad \mathbb{P}(B) = \frac{4}{52}, \qquad \mathbb{P}(B \mid A) = \frac{3}{51}.$$

Since $\mathbb{P}(B \mid A) \neq \mathbb{P}(B)$, the events are not independent.

Independence is a modeling assumption as much as a property to be checked. In many experiments it is justified by symmetry or by a mechanism designed to isolate trials from one another. In others, it is false but approximately true, which is often enough for good large-sample approximations.

2.4.1 Pairwise versus mutual independence

For more than two events, independence has layers. Pairwise independence means every pair is independent. Mutual independence requires factorization for every finite subcollection. Pairwise independence does *not* imply mutual independence.

Example 2.16 (A classic counterexample)**.** Toss two fair coins. Define the events

$$A = \{\text{first toss is heads}\}, \qquad B = \{\text{second toss is heads}\},$$

$$C = \{\text{the two tosses have the same result}\}.$$

Then any two of A, B, C are independent, but

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(HH) = \frac{1}{4} \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

So the three events are pairwise independent but not mutually independent.

This is not a technical quibble. In many theorems, especially laws of large numbers and central limit theorems, the relevant assumption is mutual independence of random variables, not just pairwise independence.

2.5 Bernoulli trials and sequential reasoning

A very large class of models begins with repeated independent trials, each with two possible outcomes, often called success and failure.

Definition 2.17. A sequence of independent trials in which each trial results in success with probability p is called a sequence of *Bernoulli trials*. The indicator of success on a single trial has a Bernoulli distribution with parameter p.

Bernoulli trials are conceptually simple but mathematically rich. They generate the binomial and geometric distributions, provide early examples of independence, and serve as the microscopic model behind many limit theorems.

Example 2.18 (Exactly k successes)**.** In n independent Bernoulli(p) trials, the probability of exactly k successes is

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

Why? Fix the locations of the k successes. The probability of that specific pattern is $p^k(1-p)^{n-k}$. There are $\binom{n}{k}$ such patterns.

This is a perfect illustration of the general strategy "count patterns, then multiply along each pattern."

2.6 Problem-solving heuristics for conditional probability

Conditional probability is a major conceptual hurdle for many students. The following habits help.

(1) **Name the information.** What exactly is being conditioned on? An event, a partition, a stage of an experiment?

(2) **Redraw the world.** After conditioning on B, pretend the sample space is B.

(3) **Use trees when the experiment is sequential.**

(4) **Use Bayes only when you are reversing information.** Do not apply it just because it is available.

(5) **Check plausibility.** If the condition makes the event more likely, the conditional probability should go up, not down.

A reliable classroom technique is to ask students to solve a problem first in words and only then with symbols. For many conditional-probability questions, a correct verbal argument is nearly equivalent to a correct formal one. The notation should clarify the argument, not intimidate it.

2.7 Summary

This chapter introduced the central mechanism of probabilistic updating.

- Conditional probability rescales the model after learning new information.
- The multiplication rule computes joint probabilities through sequential conditioning.
- The law of total probability averages over a partition of possible cases.
- Bayes' rule reverses conditions and quantifies learning from evidence.
- Independence means information about one event does not change probabilities of another.
- For more than two events, pairwise independence is weaker than mutual independence.

In the next chapter we move from events to random variables. This shift lets us summarize outcomes numerically and opens the door to probability distributions, expectation, and asymptotic theory.

Exercises

Exercise 2.1. Two cards are drawn without replacement from a standard deck. Compute the probability that both cards are hearts.

Exercise 2.2. An urn contains 4 red and 6 blue balls. Two balls are drawn without replacement. Let A be the event that the second ball is red and B the event that the first ball is red. Compute $\mathbb{P}(A)$, $\mathbb{P}(A \mid B)$, and $\mathbb{P}(A \mid B^c)$.

Exercise 2.3. A factory uses two machines. Machine 1 produces 60% of the items and machine 2 produces 40%. Machine 1 has defect rate 2% and machine 2 defect rate 5%. If a randomly chosen item is defective, what is the probability it came from machine 2?

Exercise 2.4. Suppose A and B are events with $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.5$, and $\mathbb{P}(A \cap B) = 0.2$. Are A and B independent? Compute $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$.

Exercise 2.5. Let B_1, B_2, B_3 be a partition with probabilities $0.2, 0.3, 0.5$. Suppose $\mathbb{P}(A \mid B_1) = 0.1$, $\mathbb{P}(A \mid B_2) = 0.4$, and $\mathbb{P}(A \mid B_3) = 0.7$. Compute $\mathbb{P}(A)$ and $\mathbb{P}(B_2 \mid A)$.

Exercise 2.6. Prove that if A and B are independent, then so are A and B^c, A^c and B, and A^c and B^c.

Exercise 2.7. Three fair coins are tossed. Let A be the event "an even number of heads occurs," B be the event "the first two tosses are equal," and C be the event "the last two tosses are equal." Determine which pairs among A, B, C are independent, and whether the three are mutually independent.

Exercise 2.8. In independent Bernoulli(p) trials, find the probability that the first success occurs on trial k.

Exercise 2.9. A student answers a multiple-choice question with probability 0.8 of knowing the answer. If the student knows it, the answer is correct. If not, the student guesses uniformly among 4 choices. Given that the answer is correct, what is the probability the student actually knew it?

Challenge Exercise 2.10. A family has two children. You are told that one of them is a girl born on a Tuesday. Under the usual simplifying assumptions that gender is independent and days of birth are equally likely, what is the probability both children are girls? Carefully specify your sample space before computing.

Challenge Exercise 2.11. Let $A_1, \dots, A_n$ be independent events. Show that the probability that none occurs is

$$\prod_{i=1}^{n} (1 - \mathbb{P}(A_i)).$$

Use this to derive the formula for the probability that at least one occurs.

Chapter 3

Discrete Random Variables and Their Distributions

A random variable is a numerical summary of an outcome. Once outcomes are converted into numbers, we can ask new kinds of questions: what values can occur, how likely is each value, how do we compute probabilities from a distribution, and how do different models compare?

3.1 From outcomes to random variables

Many probability questions are not really about raw outcomes. They are about a numerical quantity extracted from the outcome. In three coin tosses, we may care about the number of heads rather than the exact order. In drawing a hand of cards, we may care about the number of aces. In a queueing system, we may care about the waiting time. This motivates the next definition.

Definition 3.1. A *random variable* on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function

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

For a random variable X and a subset $B \subseteq \mathbb{R}$, the event $\{X \in B\}$ means

$$\{\omega \in \Omega : X(\omega) \in B\}.$$

In a first course, it is safe to think of a random variable simply as a function that assigns a number to each outcome. Later one refines the definition by asking that events such as $\{X \leq x\}$ belong to the underlying event collection. That technical condition is what makes probabilities like $\mathbb{P}(X \leq x)$ meaningful.

Example 3.2 (Number of heads)**.** Let $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ for three fair coin tosses. Define

$$X(\omega) = \text{number of heads in } \omega.$$

Then X takes values in $\{0, 1, 2, 3\}$. For instance,

$$\{X = 2\} = \{HHT, HTH, THH\}.$$

So

$$\mathbb{P}(X = 2) = \frac{3}{8}.$$

The notation $\{X = x\}$ is shorthand for the event that the random variable takes the specific value x. This simple notation becomes extremely efficient once the subject grows more complicated.

3.2 Probability mass functions and cumulative distribution functions

For a discrete random variable, the essential information is the probability attached to each possible value.

Definition 3.3. A random variable X is *discrete* if it takes values in a finite or countable set. Its *probability mass function* (pmf) is

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

The *support* of X is the set of points x for which $p_X(x) > 0$.

A pmf must satisfy two conditions:

(i) $p_X(x) \geq 0$ for all x;

(ii) $\sum_x p_X(x) = 1$, where the sum runs over the support.

Conversely, any function on a finite or countable set satisfying these two conditions can serve as the pmf of some discrete random variable.

Definition 3.4. The *cumulative distribution function* (cdf) of a random variable X is

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

For a discrete random variable, the cdf is a step function. It jumps exactly at the support points, and the size of the jump at x equals $\mathbb{P}(X = x)$.

Proposition 3.5. *If X is discrete, then for any real number x,*

$$F_X(x) = \sum_{t \leq x} p_X(t).$$

Moreover,

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$$

for all real numbers $a < b$.

Proof. The event $\{X \le x\}$ is the disjoint union of events $\{X = t\}$ over all support points $t \le x$. Add the probabilities. The second identity follows similarly by observing that

$$\{a < X \le b\} = \{X \le b\} \setminus \{X \le a\}.$$

□

The cdf is useful because it works for every random variable, discrete or continuous. The pmf is more specialized, but when it exists it is often more convenient for explicit calculations.

3.3 Standard discrete models

We now develop the most important discrete families. They recur throughout probability and statistics.

3.3.1 Bernoulli and binomial distributions

Definition 3.6. A random variable X has a *Bernoulli distribution* with parameter $p \in [0, 1]$, written $X \sim \mathrm{Ber}(p)$, if

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

A Bernoulli random variable records success or failure. It is the atomic building block for many later constructions.

Definition 3.7. A random variable X has a *binomial distribution* with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$, written $X \sim \mathrm{Bin}(n, p)$, if

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

The standard story is that X counts successes in n independent Bernoulli(p) trials. The formula comes from counting which k of the n positions contain successes.

Example 3.8. Suppose a multiple-choice question has four options and a student answers by guessing independently on 10 such questions. If X is the number answered correctly, then

$$X \sim \mathrm{Bin}(10, 1/4),$$

and

$$\mathbb{P}(X = 3) = \binom{10}{3} \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^7.$$

3.3.2 Geometric and negative binomial distributions

Definition 3.9. A random variable X has a *geometric distribution* with parameter $p \in (0,1)$ if

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

In this convention, X counts the trial number of the first success.

The geometric distribution describes waiting time until the first success in independent Bernoulli(p) trials. Its most distinctive property is memorylessness; we will return to this after discussing continuous waiting times.

More generally, the waiting time for the rth success has a negative binomial distribution.

Definition 3.10. A random variable X has a *negative binomial distribution* with parameters $r \in \mathbb{N}$ and $p \in (0,1)$ if

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

Here X counts the trial on which the rth success occurs.

The combinatorial factor $\binom{k-1}{r-1}$ reflects that the kth trial must be a success, and among the first $k-1$ trials exactly $r-1$ must be successes.

3.3.3 Hypergeometric distribution

The binomial model assumes independent trials, which naturally occurs with sampling with replacement or with a vast population. Without replacement, the correct model changes.

Definition 3.11. Suppose a population contains N objects, of which M are labeled success. If we sample n objects uniformly without replacement and let X be the number of successes in the sample, then X has a *hypergeometric distribution*:

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

The numerator counts favorable samples: choose k successes from the M available and $n-k$ failures from the remaining $N-M$. The denominator counts all n-element samples.

Example 3.12. A class has 18 students, 7 of whom are statistics majors. A committee of 5 is chosen uniformly at random. If X is the number of statistics majors on the committee, then

$$\mathbb{P}(X = 2) = \frac{\binom{7}{2}\binom{11}{3}}{\binom{18}{5}}.$$

3.3.4 Poisson distribution

Definition 3.13. A random variable X has a *Poisson distribution* with parameter $\lambda > 0$, written $X \sim \text{Pois}(\lambda)$, if

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \dots$$

The Poisson distribution is the canonical model for rare counts. It appears when many small independent opportunities for occurrence accumulate. Later we will see the Poisson approximation and the Poisson process, where this distribution arises dynamically over time.

A good habit is to interpret the parameter. In a Poisson model, λ measures the typical size of the count; later we will prove that $\mathbb{E}X = \lambda$ and $\text{Var}(X) = \lambda$.

3.4 Functions of a discrete random variable

Often we define a new random variable from an old one by applying a deterministic function.

Proposition 3.14. *If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function, then $Y = g(X)$ is again a random variable. For any value y,*

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

Example 3.15 (Parity of a count)**.** If $X \sim \text{Bin}(4, 1/2)$ and $Y = X \bmod 2$, then

$$\mathbb{P}(Y = 0) = \mathbb{P}(X \in \{0, 2, 4\}) = \frac{1}{16} + \frac{6}{16} + \frac{1}{16} = \frac{1}{2},$$

$$\mathbb{P}(Y = 1) = \mathbb{P}(X \in \{1, 3\}) = \frac{4}{16} + \frac{4}{16} = \frac{1}{2}.$$

So the parity is equally likely to be even or odd.

This summation-over-preimages formula is elementary but crucial. It is the discrete version of change of variables.

3.5 Jointly discrete random variables

Before developing expectation in the next chapter, it is helpful to see how several random variables can be studied together.

Definition 3.16. Two random variables X and Y are *jointly discrete* if the pair (X, Y) takes values in a finite or countable subset of $\mathbb{R}^2$. Their *joint pmf* is

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

From the joint pmf we recover the *marginal* pmfs by summing:

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

The random variables are independent exactly when

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

for all (x,y).

Example 3.17 (Two dice)**.** Let X be the first die and Y the second die when two fair dice are rolled. Then

$$p_{X,Y}(x,y) = \frac{1}{36}, \qquad x,y \in \{1,2,3,4,5,6\}.$$

Thus

$$p_X(x) = \sum_{y=1}^{6} \frac{1}{36} = \frac{1}{6}$$

for each x, and similarly for Y. Moreover,

$$p_{X,Y}(x,y) = \frac{1}{36} = \frac{1}{6}\cdot\frac{1}{6} = p_X(x)p_Y(y),$$

so X and Y are independent.

Joint distributions will be treated systematically later, but even at this stage the joint viewpoint is useful for understanding dependence and for computing distributions of sums, minima, or maxima.

3.6 Reading and using a distribution

A distribution is not just a formula. It is a compact description of how uncertainty is spread over values. When you see a new distribution, you should immediately ask:

(1) What does the variable count or measure?

(2) What are the possible values?

(3) Which parameters control location, spread, or shape?

(4) Is there a natural story that generates the model?

(5) Which approximations or related models does it connect to?

For example:

- Bernoulli models one trial.
- Binomial models a fixed number of independent trials.

- Geometric models waiting time to the first success.
- Hypergeometric models sampling without replacement.
- Poisson models rare counts.

These stories matter because they guide model selection. Much of applied probability is choosing the right distributional story before computing anything.

3.7 Summary

This chapter developed the language of discrete random variables.

- A random variable is a numerical summary of an outcome.
- A discrete random variable is described by its pmf, and equivalently by its cdf.
- The standard discrete families—Bernoulli, binomial, geometric, negative binomial, hypergeometric, and Poisson—each arise from a characteristic sampling story.
- Functions of a random variable are analyzed by summing probabilities over preimages.
- Joint pmfs capture dependence between random variables.

The next chapter studies expectation and variance, the two most important numerical summaries of a distribution.

Exercises

Exercise 3.1. Three fair coins are tossed. Let X be the number of heads. Write down the pmf and cdf of X.

Exercise 3.2. If $X \sim \text{Bin}(8, 0.3)$, compute $\mathbb{P}(X = 2)$, $\mathbb{P}(X \leq 2)$, and $\mathbb{P}(X \geq 1)$.

Exercise 3.3. A basketball player makes each free throw independently with probability 0.8. Let X be the number of shots until the first miss. Identify the distribution of X and compute $\mathbb{P}(X = 5)$.

Exercise 3.4. An urn contains 10 balls, 4 of which are red. Five balls are drawn without replacement. Let X be the number of red balls drawn. Compute $\mathbb{P}(X = 2)$ and $\mathbb{P}(X \geq 1)$.

Exercise 3.5. If $X \sim \text{Pois}(3)$, compute $\mathbb{P}(X = 0)$, $\mathbb{P}(X = 1)$, and $\mathbb{P}(X \leq 2)$.

Exercise 3.6. Let X have pmf $p(0) = 1/6$, $p(1) = 1/3$, $p(2) = 1/2$. Let $Y = X^2$. Find the pmf of Y.

Exercise 3.7. Suppose X and Y have joint pmf

$$p_{X,Y}(x,y) = c(x+y), \qquad x,y \in \{0,1,2\}.$$

Find the constant c, the marginal pmfs, and determine whether X and Y are independent.

Exercise 3.8. Prove that for any discrete random variable X,

$$\sum_x p_X(x) = 1.$$

Then explain why the support must be at most countable.

Challenge Exercise 3.9. Show that if X is geometric with parameter p, then for all integers $m, n \geq 1$,

$$\mathbb{P}(X > m+n \mid X > m) = \mathbb{P}(X > n).$$

Interpret this identity in words.

Challenge Exercise 3.10. Let $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ be independent. Guess the distribution of $X+Y$ and prove your guess directly from the pmf using a combinatorial identity.

Chapter 4

Expectation, Variance, and Basic Inequalities

A probability distribution tells us what can happen; expectation and variance summarize how the distribution is centered and how widely it spreads. These summaries are not merely descriptive. They are also computational tools, especially because expectation is linear and variance interacts cleanly with sums of independent random variables.

4.1 Expectation as a weighted average

For a discrete random variable, the expectation is the probability-weighted average of its values.

Definition 4.1. If X is a discrete random variable taking values $x_1, x_2, \dots$ with pmf p_X, and if the sum converges absolutely, the *expectation* of X is

$$\mathbb{E}[X] = \sum_x x\, p_X(x).$$

The phrase "converges absolutely" matters. Without it, positive and negative parts could separately diverge, making the weighted average ill-defined. In elementary examples, this subtlety is often invisible, but it becomes important in serious probability.

Expectation should be understood in three complementary ways.

(1) It is a long-run average in repeated sampling.

(2) It is a center of mass of the distribution.

(3) It is a linear operator that converts random variables into numbers.

All three viewpoints are useful. The first gives intuition, the second gives geometric meaning, and the third powers many calculations.

Example 4.2. If $X \sim \mathrm{Ber}(p)$, then

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

If $X \sim \mathrm{Bin}(n,p)$ and we already know X counts successes in n Bernoulli trials, we may guess that $\mathbb{E}[X] = np$. We will prove this shortly in a way that scales to many other problems.

4.1.1 Linearity of expectation

The single most important property of expectation is linearity.

Theorem 4.3 (Linearity)**.** *If X and Y are integrable random variables and $a, b \in \mathbb{R}$, then*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

More generally,

$$\mathbb{E}\Big[\sum_{i=1}^{n} a_i X_i\Big] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i].$$

Proof. For discrete random variables, the statement follows from distributing the sum and using ordinary algebra. In the finite-support case the proof is immediate; the general discrete case follows from absolute convergence. □

A remarkable feature of linearity is that it does *not* require independence. This cannot be emphasized enough. Expectation interacts with sums in a universally simple way, even when the summands are dependent.

Example 4.4 (Expected number of heads)**.** Let X be the number of heads in n independent Bernoulli(p) trials. Write

$$X = I_1 + \cdots + I_n,$$

where I_j is the indicator that trial j is a success. Then each I_j has expectation p, so

$$\mathbb{E}[X] = \mathbb{E}[I_1] + \cdots + \mathbb{E}[I_n] = np.$$

This argument is shorter and more informative than computing the binomial sum directly.

4.2 The law of the unconscious statistician

Often we need the expectation of a function of a random variable.

Proposition 4.5 (LOTUS)**.** *If X is discrete and g is a function for which the sum converges absolutely, then*

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

Students often rediscover this formula from scratch by first finding the distribution of $g(X)$. That approach works, but it can be unnecessarily laborious. LOTUS allows us to compute directly from the distribution of X itself.

Example 4.6. If $X \sim \mathrm{Geom}(p)$, then

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

Evaluating this series gives $1/p$. A slick way is to start from

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}, \qquad |q| < 1,$$

and differentiate with respect to q.

4.3 Indicators and counting arguments

Indicator random variables deserve a chapter of their own, but for now we develop the essential trick.

Definition 4.7. For an event A, the *indicator* of A is the random variable

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

Its expectation is especially simple:

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

The proof is immediate from the pmf of a Bernoulli random variable.

The power of indicators comes from decomposition. If X counts how many of several things happen, we can often write

$$X = \sum_{i=1}^{n} \mathbf{1}_{A_i}$$

for suitable events A_i. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

This avoids the need to find the entire distribution of X.

Example 4.8 (Fixed points in a random permutation)**.** Let π be a uniformly random permutation of $\{1, 2, \dots, n\}$. Let X be the number of fixed points, that is, positions i such that $\pi(i) = i$. Define $I_i = \mathbf{1}_{\{\pi(i)=i\}}$. Then

$$X = \sum_{i=1}^{n} I_i, \qquad \mathbb{E}[I_i] = \frac{1}{n}.$$

Therefore

$$\mathbb{E}[X] = \sum_{i=1}^{n} \frac{1}{n} = 1.$$

Notice how little work this required. We did not need the full distribution of X.

Example 4.9 (Occupancy problem)**.** Throw m balls independently into n boxes, each ball equally likely to land in any box. Let X be the number of empty boxes. For box i, let I_i be the indicator that box i is empty. Then

$$\mathbb{E}[I_i] = \left(1 - \frac{1}{n}\right)^m,$$

so

$$\mathbb{E}[X] = n\left(1 - \frac{1}{n}\right)^m.$$

Again, the whole distribution would be difficult, but the expectation is easy.

4.4 Variance and spread

Expectation tells us where a distribution is centered; it does not tell us how spread out the values are. Variance addresses this.

Definition 4.10. If X has finite second moment, the *variance* of X is

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

The square root of the variance is the *standard deviation*.

Squaring ensures that deviations above and below the mean do not cancel. It also heavily penalizes large deviations, which is both a strength and a limitation.

Proposition 4.11 (Useful formula)**.** *If X has finite second moment, then*

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

Proof. Expand the square:

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

Take expectations and use linearity. □

Example 4.12. If $X \sim \operatorname{Ber}(p)$, then $X^2 = X$, so

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

If $X = I_1 + \cdots + I_n$ is binomial with independent Bernoulli(p) indicators, then one can show

$$\operatorname{Var}(X) = np(1-p).$$

We will derive this from covariance shortly.

4.4.1 Effect of affine transformations

Proposition 4.13. *For constants $a, b \in \mathbb{R}$,*

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

Adding a constant shifts the distribution but does not change spread. Multiplying by a rescales deviations by a, so variance is multiplied by a^2.

4.5 Covariance and sums of random variables

When two random variables vary together, covariance captures their linear association.

Definition 4.14. If X and Y have finite second moments, their *covariance* is

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

Equivalent algebra gives

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, hence $\mathrm{Cov}(X, Y) = 0$. The converse is false in general: zero covariance need not imply independence.

Theorem 4.15 (Variance of a sum)**.** *If X and Y have finite second moments, then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

More generally,

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j).$$

Corollary 4.16. *If $X_1, \ldots, X_n$ are pairwise uncorrelated—in particular, if they are independent—then*

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Example 4.17 (Binomial variance)**.** If $X = I_1 + \cdots + I_n$ with independent Bernoulli(p) indicators, then

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

This is much simpler than trying to compute $\mathbb{E}[X^2]$ directly from the binomial pmf.

4.6 Tail-sum formula and related identities

Expectation of a nonnegative integer-valued random variable has a useful alternative representation.

Theorem 4.18 (Tail-sum formula)**.** *If X is a nonnegative integer-valued random variable, then*

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

Proof. Write

$$X = \sum_{k=1}^{\infty} \mathbf{1}_{\{X \geq k\}}.$$

Take expectations and use linearity. □

This formula is often the right tool for waiting-time problems.

Example 4.19 (Coupon collector: first step)**.** Suppose one repeatedly samples uniformly from n coupon types. Let T be the number of draws needed to see the first coupon type. Trivial. But for more complicated waiting times, tail probabilities are often easier to compute than pmfs. The tail-sum formula turns those tail computations into expectations.

A better example is the geometric distribution. If $X \sim \text{Geom}(p)$, then

$$\mathbb{P}(X \geq k) = (1-p)^{k-1},$$

so

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

This derivation is both short and conceptually clean.

4.7 Basic inequalities

Probability theory is full of identities, but it is also full of inequalities. They let us control complicated quantities by simpler ones.

4.7.1 Markov's inequality

Theorem 4.20 (Markov)**.** *If $X \geq 0$ and $a > 0$, then*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Since $X \geq a\mathbf{1}_{\{X \geq a\}}$, taking expectations yields

$$\mathbb{E}[X] \geq a\mathbb{P}(X \geq a).$$

Rearrange. □

Markov's inequality is crude but very general. It bounds a tail probability using only the mean.

4.7.2 Chebyshev's inequality

Theorem 4.21 (Chebyshev)**.** *If X has finite variance and $t > 0$, then*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. Apply Markov's inequality to the nonnegative random variable $(X - \mathbb{E}[X])^2$. □

Chebyshev's inequality is one of the first rigorous ways to convert variance information into concentration around the mean. It is the engine behind the weak law of large numbers.

4.7.3 Cauchy–Schwarz

Theorem 4.22 (Cauchy–Schwarz inequality)**.** *If X and Y have finite second moments, then*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.$$

Proof. For any real t,

$$0 \leq \mathbb{E}[(X + tY)^2] = \mathbb{E}[X^2] + 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2].$$

The quadratic in t must have discriminant at most 0, which gives the result. □

An important corollary is the bound

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

This motivates the next normalized measure of linear association.

Definition 4.23. If $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$, the *correlation coefficient* is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

By Cauchy–Schwarz, $-1 \leq \mathrm{Corr}(X, Y) \leq 1$.

4.8 Expected value as a modeling tool

At this point it is worth stepping back. Expectation is not just a number you compute after the model is finished. It is often part of the modeling language itself.

When an insurer prices risk, it begins with expected payout. When a queueing theorist estimates congestion, expected waiting time is a central quantity. When a statistician studies an estimator, expectation measures bias and variance measures precision. These are not separate worlds; they are different uses of the same mathematical operator.

The indicator method is especially important in applications because it turns global counts into local probabilities. Whenever you are asked to compute the expected number of something, ask yourself whether that "something" can be decomposed into contributions from smaller pieces.

4.9 Summary

This chapter developed the core numerical summaries of distributions.

- Expectation is a weighted average and a linear operator.
- Indicator random variables convert counting questions into expectation problems.
- Variance measures spread, and covariance measures linear co-movement.
- Variances add cleanly for independent sums.
- Markov, Chebyshev, and Cauchy–Schwarz provide the first general probability bounds.

In the next chapter we move from discrete models to continuous random variables, where probabilities are described by densities and integrals replace sums.

Exercises

Exercise 4.1. Compute the expectation and variance of a Bernoulli(p) random variable.

Exercise 4.2. Let $X \sim \mathrm{Bin}(n, p)$. Write X as a sum of indicators and use this representation to derive $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

Exercise 4.3. If $X \sim \mathrm{Geom}(p)$, use the tail-sum formula to compute $\mathbb{E}[X]$.

Exercise 4.4. Let X be the number of red balls in a sample of size n drawn without replacement from a population of size N containing M red balls. Use indicators to compute $\mathbb{E}[X]$.

Exercise 4.5. A fair die is rolled 12 times. Let X be the number of sixes. Compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and use Chebyshev's inequality to bound $\mathbb{P}(|X - 2| \geq 2)$.

Exercise 4.6. Suppose X is nonnegative and $\mathbb{E}[X] = 10$. What can Markov's inequality tell you about $\mathbb{P}(X \geq 25)$? Give an example where the bound is attained.

Exercise 4.7. Let X and Y have means 0, variances 4 and 9, and covariance -3. Compute $\mathrm{Var}(X + Y)$ and $\mathrm{Var}(X - Y)$.

Exercise 4.8. Prove that if X is integer-valued and nonnegative, then

$$\mathbb{E}[X(X-1)] = \sum_{k=2}^{\infty} k(k-1)\mathbb{P}(X = k).$$

Then evaluate this quantity when $X \sim \mathrm{Bin}(n, p)$.

Challenge Exercise 4.9. A random permutation of $\{1, 2, \ldots, n\}$ is chosen uniformly. Let X be the number of fixed points. Compute $\mathrm{Var}(X)$. Hint: write X as a sum of indicators and evaluate the pairwise covariances.

Challenge Exercise 4.10. Let X be a nonnegative integer-valued random variable. Show that

$$\mathbb{E}[X^2] = \sum_{k=1}^{\infty} (2k-1)\mathbb{P}(X \geq k).$$

Use this identity to compute $\mathbb{E}[X^2]$ when $X \sim \mathrm{Geom}(p)$.

Chapter 5

Continuous Random Variables

For a continuous random variable, probabilities are not attached to individual points but to intervals and regions. Densities describe how probability is distributed continuously across the line, and integrals replace sums.

5.1 From mass functions to densities

Discrete random variables are described by point masses $\mathbb{P}(X = x)$. Continuous models behave differently. If a measurement is recorded on a genuinely continuous scale, the chance of observing any one exact value is typically 0. What matters is probability over intervals.

Definition 5.1. A random variable X is *continuous with density* f if there exists a nonnegative function f on $\mathbb{R}$ such that

$$\mathbb{P}(a \le X \le b) = \int_a^b f(x)\,dx$$

for all real numbers $a \le b$.

The function f is called the *probability density function* (pdf) of X. It is not itself a probability. Probabilities are areas under the density curve.

A valid density must satisfy

$$f(x) \ge 0 \quad \text{for all } x, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Conversely, any function satisfying these two conditions is a density.

Remark 5.2. If X has a density, then for every point x,

$$\mathbb{P}(X = x) = 0.$$

This is not paradoxical. A single point has zero width, so the corresponding area under the density is zero. Nevertheless, intervals of nonzero width can have substantial probability.

5.2 Distribution functions and densities

As before, the cdf is defined by

$$F_X(x) = \mathbb{P}(X \le x).$$

If X has density f, then

$$F_X(x) = \int_{-\infty}^{x} f(t)\,dt.$$

Thus the cdf is the accumulated area under the density curve up to x.

If f is continuous, then the Fundamental Theorem of Calculus gives

$$F_X'(x) = f(x).$$

So densities and cdfs are two views of the same object: the density is the derivative of the cdf, while the cdf is the integral of the density.

Example 5.3 (Uniform density on an interval)**.** If $X \sim \text{Unif}(a, b)$, then

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b,$$

and $f(x) = 0$ otherwise. The cdf is

$$F_X(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

The distribution is uniform because equal-length subintervals carry equal probability.

5.3 Computing probabilities from a density

Once a density is known, interval probabilities are found by integration.

- $\mathbb{P}(X \le x) = \int_{-\infty}^{x} f(t)\,dt.$
- $\mathbb{P}(a < X \le b) = \int_{a}^{b} f(t)\,dt.$
- $\mathbb{P}(X > x) = 1 - F_X(x).$

Because point probabilities are zero, the choice of strict or non-strict inequalities is irrelevant for a continuous random variable:

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

Example 5.4. Let X have density

$$f(x) = 2x, \qquad 0 \le x \le 1,$$

and $f(x) = 0$ otherwise. Then

$$F_X(x) = x^2 \quad \text{for } 0 \le x \le 1.$$

So

$$\mathbb{P}\left(\frac{1}{2} \le X \le 1\right) = 1 - \left(\frac{1}{2}\right)^2 = \frac{3}{4}.$$

5.4 Expectation and variance in the continuous case

Sums become integrals.

Definition 5.5. If X has density f and the integral converges absolutely, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

More generally,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

Variance is defined exactly as before:

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

provided the relevant integrals exist.

Example 5.6 (Uniform distribution)**.** If $X \sim \text{Unif}(a, b)$, then

$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a}\, dx = \frac{a+b}{2}.$$

Similarly,

$$\mathbb{E}[X^2] = \frac{1}{b-a} \int_a^b x^2\, dx = \frac{a^2 + ab + b^2}{3},$$

which yields

$$\operatorname{Var}(X) = \frac{(b-a)^2}{12}.$$

5.5 Important continuous families

We now develop the standard continuous models that appear repeatedly in applications.

5.5.1 Uniform distribution

The uniform distribution on an interval is the continuous analog of a finite uniform model. It describes ignorance over a bounded range when no point in the interval is favored over another. It is also useful as a simulation engine because many other distributions can be constructed from it.

5.5.2 Exponential distribution

Definition 5.7. A random variable X has an *exponential distribution* with rate $\lambda > 0$, written $X \sim \mathrm{Exp}(\lambda)$, if

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0,$$

and $f(x) = 0$ for $x < 0$.

Its cdf is

$$F_X(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

A short integration gives

$$\mathbb{E}[X] = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

The exponential distribution is the canonical continuous waiting-time model. It plays a role parallel to the geometric distribution in discrete time.

Theorem 5.8 (Memoryless property)**.** *If $X \sim \mathrm{Exp}(\lambda)$, then for all $s, t \geq 0$,*

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t).$$

Proof. Since $\mathbb{P}(X > u) = e^{-\lambda u}$,

$$\mathbb{P}(X > s + t \mid X > s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t).$$

□

This says that, conditional on having already waited time s, the additional waiting time has the same distribution as the original waiting time. In many stochastic-process models, this property is a signature of complete absence of aging.

5.5.3 Normal distribution

Definition 5.9. A random variable X has a *normal distribution* with mean μ and variance $\sigma^2 > 0$, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

The standard normal distribution is $\mathcal{N}(0,1)$; its cdf is often denoted by Φ. Unlike the uniform and exponential cases, the normal cdf does not have an elementary antiderivative, so probabilities are usually computed numerically or from tables. Nevertheless, the normal family is central because of the central limit theorem.

Affine transformations behave cleanly:

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

This is called *standardization*.

Remark 5.10. The normal distribution is mathematically special for several reasons. Sums of independent normal random variables remain normal, the density is infinitely differentiable, and the quadratic exponent makes it deeply compatible with Fourier methods and asymptotic approximations.

5.5.4 Gamma distribution

The gamma family generalizes the exponential distribution.

Definition 5.11. A random variable X has a *gamma distribution* with shape $\alpha > 0$ and rate $\lambda > 0$ if its density is

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0,$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx.$$

When $\alpha = 1$, this reduces to the exponential distribution. Later, in the Poisson-process chapter, we will see that waiting time until the kth arrival has a gamma distribution with integer shape k.

5.5.5 Beta distribution

Definition 5.12. A random variable X has a *beta distribution* with parameters $\alpha, \beta > 0$ if its density is

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1,$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx.$$

The beta family is flexible on the interval $(0,1)$ and often models proportions or probabilities. Depending on the parameters, it can be symmetric, skewed, U-shaped, or mound-shaped.

5.6 Quantiles, medians, and percentiles

The mean is not the only measure of location. Another important summary comes from the cdf.

Definition 5.13. A number m is a *median* of X if

$$\mathbb{P}(X \le m) \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \ge m) \ge \frac{1}{2}.$$

More generally, a *p-quantile* is a value q_p satisfying roughly

$$\mathbb{P}(X \le q_p) = p.$$

For a strictly increasing cdf, the p-quantile is just $F^{-1}(p)$. Quantiles are often more robust than means when distributions are skewed or heavy-tailed.

Example 5.14. If $X \sim \text{Exp}(\lambda)$, the median m solves

$$1 - e^{-\lambda m} = \frac{1}{2},$$

so

$$m = \frac{\log 2}{\lambda}.$$

Notice that the median is smaller than the mean $1/\lambda$, reflecting the right-skew of the exponential distribution.

5.7 Mixed and non-continuous distributions

Not every random variable is purely discrete or purely continuous. One can have a mixed distribution: some probability mass at specific points together with a density on an interval.

Example 5.15. Suppose a machine fails immediately with probability 0.1, and otherwise survives for an exponential amount of time with rate λ. Then the lifetime T satisfies

$$\mathbb{P}(T = 0) = 0.1,$$

while for $t > 0$ the remaining probability is spread continuously. The cdf has a jump at 0 and is smooth afterward.

This reminds us not to overgeneralize. The rule "point probabilities are zero" is true for continuous random variables with densities, not for all random variables.

5.8 How densities should be interpreted

A density is a local concentration of probability, not a probability itself. When $f(x)$ is large, nearby values are relatively likely compared with nearby regions where the density is small. A useful approximation is

$$\mathbb{P}(x \le X \le x + h) \approx f(x)h$$

for small $h > 0$ when f is continuous at x.

This approximation explains why densities can exceed 1. For instance, the uniform density on $(0, 0.1)$ is 10. That does not violate any rule, because probability comes from area, and the total area is still 1.

5.9 Summary

This chapter developed the calculus-based side of probability.

- Continuous random variables are described by densities, and probabilities are computed by integration.
- The cdf accumulates area under the density and differentiates back to the density when smoothness holds.
- Expectation and variance in the continuous case are computed by integrals.
- The main continuous families for this course are the uniform, exponential, normal, gamma, and beta distributions.
- The exponential distribution is memoryless and serves as the basic waiting-time model in continuous time.

The next chapter studies several random variables together: joint densities, marginals, conditionals, and independence.

Exercises

Exercise 5.1. Show that the function

$$f(x) = c(1 - x^2), \qquad -1 \le x \le 1,$$

with $f(x) = 0$ otherwise, is a density for an appropriate choice of c. Find c.

Exercise 5.2. Let X have density $f(x) = 2x$ on $[0, 1]$. Compute the cdf, the mean, and the variance of X.

Exercise 5.3. If $X \sim \text{Unif}(-2, 4)$, compute $\mathbb{P}(X > 1)$, $\mathbb{E}[X]$, and $\text{Var}(X)$.

Exercise 5.4. If $X \sim \text{Exp}(\lambda)$, compute $\mathbb{P}(X > t)$ and verify the memoryless property.

Exercise 5.5. Suppose X has density $f(x) = \frac{1}{2}e^{-|x|}$ on $\mathbb{R}$. Verify that f is a density and compute $\mathbb{P}(|X| \le 1)$.

Exercise 5.6. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Show that $Y = (X - \mu)/\sigma$ has density

$$\frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Exercise 5.7. Find the median of a uniform distribution on (a, b).

Exercise 5.8. A lifetime has density $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. Find the 90th percentile.

Challenge Exercise 5.9. Show that among all intervals of fixed length $2a$, the interval centered at 0 has the largest probability under the standard normal distribution.

Challenge Exercise 5.10. Suppose X has density f and finite mean. Show that if f is symmetric about 0, that is $f(x) = f(-x)$ for all x, then $\mathbb{E}[X] = 0$.

Chapter 6

Joint Distributions, Marginals, and Conditioning

Real probabilistic systems usually involve several random quantities at once. Joint distributions describe how random variables interact; marginals describe the behavior of each variable alone; conditional distributions describe what remains uncertain after partial information is revealed.

6.1 Why study random variables together?

A single random variable can describe a count, a waiting time, or a measurement. But many interesting questions involve relationships between variables: the number of heads on the first half and second half of a sequence of tosses; the lifetime of two components in a system; the height and weight of a person; the waiting time for the next bus and the total commute time. To study such questions, we need joint distributions.

If X and Y are random variables defined on the same experiment, then the pair (X, Y) is itself a random object taking values in the plane. The probability model for (X, Y) records not only the separate behavior of X and Y, but also how they depend on one another.

6.2 Joint pmfs and joint densities

6.2.1 Discrete case

If (X, Y) is jointly discrete, its joint pmf is

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

The joint pmf must satisfy

$$p_{X,Y}(x,y) \geq 0, \qquad \sum_x \sum_y p_{X,Y}(x,y) = 1.$$

Marginal distributions are recovered by summing over the other variable:

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

6.2.2 Continuous case

If (X,Y) is continuous with joint density $f_{X,Y}$, then

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dx\,dy$$

for measurable regions $A \subseteq \mathbb{R}^2$.

The marginal densities are obtained by integrating out the other coordinate:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

Geometrically, the joint density spreads probability over the plane; marginalization projects that probability onto one axis.

Example 6.1 (Uniform distribution on a triangle)**.** Suppose (X,Y) is uniformly distributed on the triangular region

$$T = \{(x,y) : 0 < y < x < 1\}.$$

The area of T is $1/2$, so the joint density is

$$f_{X,Y}(x,y) = 2, \qquad 0 < y < x < 1.$$

Then

$$f_X(x) = \int_0^x 2\,dy = 2x, \qquad 0 < x < 1,$$

and

$$f_Y(y) = \int_y^1 2\,dx = 2(1-y), \qquad 0 < y < 1.$$

Even though the joint density is constant on the triangle, the marginals are not uniform.

This example is an important warning: simple joint geometry does not necessarily translate into simple one-dimensional behavior.

6.3 Independence of random variables

The idea of independence extends from events to random variables.

Definition 6.2. Random variables X and Y are *independent* if for all Borel sets $A, B \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

In the discrete case this is equivalent to

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

for all (x, y). In the continuous density case it is equivalent to

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

for all (x, y).

The factorization criterion is extremely convenient. It lets us check independence algebraically from the joint distribution.

Example 6.3 (Independent exponentials)**.** Suppose X and Y have joint density

$$f_{X,Y}(x, y) = \lambda\mu e^{-\lambda x - \mu y}, \qquad x, y > 0.$$

Then

$$f_{X,Y}(x, y) = \big(\lambda e^{-\lambda x}\big)\big(\mu e^{-\mu y}\big) = f_X(x)f_Y(y),$$

so X and Y are independent, with $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$.

Remark 6.4. Independence is a property of the joint distribution, not of the marginals alone. Two random variables can have exactly the same marginal distributions and yet be highly dependent.

6.4 Conditional distributions

The distribution of one variable may change when the value of another is known. This is the distributional version of conditional probability.

6.4.1 Discrete conditional distributions

If X and Y are jointly discrete and $\mathbb{P}(Y = y) > 0$, define

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

For fixed y, this is a pmf in x.

Example 6.5 (Balls in two boxes)**.** Suppose two balls are independently placed into box 1 with probability p and box 2 with probability $1-p$. Let X be the number of balls in box 1 and let Y indicate whether the first ball landed in box 1. Given $Y = 1$, the first ball is already in box 1, so the conditional distribution of X is shifted compared with its unconditional distribution. Conditional distributions capture such changes precisely.

6.4.2 Continuous conditional densities

If (X, Y) has a joint density and $f_Y(y) > 0$, define the *conditional density* of X given $Y = y$ by

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Then for each fixed y, the function $x \mapsto f_{X|Y}(x \mid y)$ integrates to 1.

Example 6.6 (Uniform on a triangle, revisited)**.** For the triangular density above,

$$f_{X,Y}(x, y) = 2, \qquad 0 < y < x < 1,$$

and

$$f_Y(y) = 2(1-y), \qquad 0 < y < 1.$$

Therefore

$$f_{X|Y}(x \mid y) = \frac{2}{2(1-y)} = \frac{1}{1-y}, \qquad y < x < 1.$$

So conditional on $Y = y$, the variable X is uniformly distributed on $(y, 1)$.

This is a good illustration of how conditioning changes the support as well as the formula.

6.5 The law of total expectation and iterated conditioning

Conditional distributions make it possible to calculate probabilities and expectations in stages. One of the fundamental identities is the law of total expectation, also called the tower property in its simplest form.

Theorem 6.7 (Law of total expectation: discrete conditioning)**.** *If X and Y are discrete and X is integrable, then*

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y).$$

Proof. Starting from the definition of conditional expectation in the discrete setting,

$$\sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) = \sum_y \sum_x x\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y).$$

Since

$$\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) = \mathbb{P}(X = x, Y = y),$$

this becomes

$$\sum_y \sum_x x\mathbb{P}(X = x, Y = y) = \sum_x x \sum_y \mathbb{P}(X = x, Y = y) = \sum_x x\mathbb{P}(X = x) = \mathbb{E}[X].$$

□

The continuous version has the same spirit but uses integrals. One can think of it as averaging conditional means against the conditioning variable.

Example 6.8 (Random sum with random parameter)**.** A machine produces a Poisson number N of defects in a day, but the rate depends on the shift: on day shift the rate is 2, on night shift the rate is 5. Suppose each shift is equally likely. Then

$$\mathbb{E}[N] = \mathbb{E}[\mathbb{E}[N \mid \text{shift}]] = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 5 = 3.5.$$

The inner expectation is easy because once the shift is known, the model is simple.

6.6 Expectation of functions of two variables

For jointly distributed variables, the analog of LOTUS is straightforward.

Proposition 6.9. *If (X, Y) is jointly discrete and g is a function such that the sum converges absolutely, then*

$$\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y)p_{X,Y}(x, y).$$

If (X, Y) has joint density $f_{X,Y}$ and the integral is absolutely convergent, then

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

Example 6.10 (Product moments)**.** If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y],$$

provided the expectations exist. In the density case this follows because

$$\iint xyf_X(x)f_Y(y)\,dx\,dy = \left(\int xf_X(x)\,dx\right)\left(\int yf_Y(y)\,dy\right).$$

This identity is not merely computational. It is one of the most useful signatures of independence.

6.7 Covariance and correlation revisited

Now that we have joint distributions, covariance becomes easier to interpret.

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

If X tends to be large when Y is large, covariance is usually positive. If one tends to be large when the other is small, covariance is usually negative. If linear tendencies cancel out, covariance can be zero even when dependence remains.

Example 6.11 (Zero covariance without independence)**.** Let X take values $-1, 0, 1$ each with probability $1/3$, and let $Y = X^2$. Then Y is completely determined by X, so the variables are certainly not independent. But

$$\mathbb{E}[X] = 0, \qquad \mathbb{E}[XY] = \mathbb{E}[X^3] = 0,$$

so

$$\mathrm{Cov}(X, Y) = 0.$$

Dependence can exist without linear association.

Correlation normalizes covariance to the scale $[-1, 1]$, but it still measures only linear dependence. This is a strength and a weakness: easy to interpret, but incomplete.

6.8 Conditional probability as a geometric operation

In two dimensions, conditional distributions can often be visualized clearly. Imagine a joint density over the plane. To condition on $Y = y$, freeze a horizontal slice at height y, then renormalize the remaining function in the x-direction so that the total area becomes 1. Marginalization, by contrast, collapses the full density by integrating in the perpendicular direction.

This geometric perspective is especially helpful in continuous problems where students can become buried in notation. The formulas are important, but the picture often reveals the logic more quickly.

6.9 The bivariate normal as a preview

A full study of the multivariate normal distribution is beyond the needs of the present course, but it is worth a brief look because it illustrates how joint distributions encode both marginal behavior and dependence.

A pair (X, Y) is jointly normal if its joint density has the form

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}Q(x, y)\right),$$

where $Q(x, y)$ is an appropriate quadratic form. The parameter ρ controls linear dependence. When $\rho = 0$, the density factorizes and the variables are independent. Unlike most families, the normal distribution has the remarkable property that uncorrelated jointly normal variables are automatically independent.

We mention this now because the multivariate normal will reappear naturally when we study the central limit theorem.

6.10 Summary

This chapter developed the basic calculus of several random variables.

- Joint distributions describe the simultaneous behavior of random variables.
- Marginals are obtained by summing or integrating out other variables.
- Independence corresponds to factorization of the joint distribution.
- Conditional distributions describe how uncertainty changes after partial information is known.
- Expectation can often be computed in stages by conditioning.
- Covariance and correlation summarize linear dependence but do not capture all dependence.

The next chapter studies transformations of random variables, including change of variables, sums, minima, maxima, and order statistics.

Exercises

Exercise 6.1. Let (X, Y) have joint pmf

$$p(x, y) = cxy, \qquad x \in \{1, 2\},\ y \in \{1, 2, 3\}.$$

Find c, the marginals, and determine whether X and Y are independent.

Exercise 6.2. Suppose (X, Y) is uniformly distributed on the rectangle $0 < x < 2$, $0 < y < 1$. Find the joint density and the marginal densities. Are X and Y independent?

Exercise 6.3. Let (X, Y) have joint density

$$f(x, y) = x + y, \qquad 0 < x < 1,\ 0 < y < 1,$$

normalized appropriately. Find the normalization constant, the marginal densities, and $\mathbb{P}(X < Y)$.

Exercise 6.4. For the triangular density from the chapter, compute $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{P}(X > 1/2 \mid Y = 1/4)$.

Exercise 6.5. Show that if X and Y are independent and integrable, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Exercise 6.6. Give an example of dependent random variables with the same marginal distributions as two independent fair coin indicators.

Exercise 6.7. Let X and Y be jointly discrete. Prove the law of total expectation

$$\mathbb{E}[X] = \sum_{y} \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y).$$

Exercise 6.8. Suppose $X \sim \mathrm{Ber}(1/2)$ and $Y = X$. Compute $\mathrm{Cov}(X, Y)$ and $\mathrm{Corr}(X, Y)$.

Challenge Exercise 6.9. Let X be uniformly distributed on $(-1, 1)$ and let $Y = X^2$. Compute $\mathrm{Cov}(X, Y)$ and explain why the answer does not contradict dependence.

Challenge Exercise 6.10. Suppose X and Y are independent exponential random variables with rates λ and μ. Find the probability that $X < Y$.

Chapter 7

Transformations, Convolution, and Order Statistics

Once a distribution is known, many new random variables can be built from it by deterministic transformation: sums, differences, products, minima, maxima, and ranks. The task of this chapter is to learn how distributions change under such operations.

7.1 Why transformations matter

In applications we rarely observe random variables in isolation. We add them, scale them, apply nonlinear functions, and compare them. A statistician studies a sample mean, not merely the individual observations. A reliability engineer studies the minimum of component lifetimes. A data analyst studies the maximum of a sample or the range. A queueing theorist studies the sum of service times.

The central question is always the same: if we know the distribution of X (or of several variables), what can we say about the distribution of some new variable such as $g(X)$ or $X + Y$?

7.2 One-dimensional transformations

7.2.1 The cdf method

The most robust general technique is the cdf method. Let $Y = g(X)$. Then for any y,

$$F_Y(y) = \mathbb{P}(g(X) \le y).$$

If one can rewrite the event $\{g(X) \le y\}$ in terms of X, then the cdf of Y follows.

Example 7.1 (Square of a uniform variable)**.** Let $X \sim \mathrm{Unif}(0, 1)$ and define $Y = X^2$. Then for

$0 \le y \le 1$,

$$F_Y(y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(X \le \sqrt{y}) = \sqrt{y}.$$

Differentiating gives

$$f_Y(y) = \frac{1}{2\sqrt{y}}, \qquad 0 < y < 1.$$

The cdf method is reliable even when g is not one-to-one. Its downside is that the algebra can become cumbersome.

7.2.2 Monotone change of variables

When g is monotone and differentiable, there is a faster density formula.

Theorem 7.2 (One-dimensional change of variables)**.** *Suppose X has density f_X and $Y = g(X)$ where g is differentiable and strictly monotone with inverse $h = g^{-1}$. Then*

$$f_Y(y) = f_X(h(y))\,|h'(y)|,$$

for y in the range of g.

Proof. Assume first that g is increasing. Then

$$F_Y(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \le h(y)) = F_X(h(y)).$$

Differentiate with respect to y and apply the chain rule. The decreasing case is similar and introduces the absolute value. □

Example 7.3 (Affine transformation)**.** Let $Y = aX + b$ with $a \neq 0$. Then $h(y) = (y - b)/a$, so

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

This general formula is worth memorizing. It explains how location and scale changes affect densities.

Example 7.4 (Exponential from a uniform)**.** If $U \sim \mathrm{Unif}(0,1)$ and

$$X = -\frac{1}{\lambda}\log U,$$

then $X \sim \mathrm{Exp}(\lambda)$. Indeed, $U = e^{-\lambda X}$, so

$$f_X(x) = 1 \cdot \lambda e^{-\lambda x}, \qquad x > 0.$$

This gives the inverse-cdf method for simulating exponential random variables from a uniform source.

7.3 Many-to-one transformations

When g is not one-to-one, several points of the original variable may map to the same value. Then we must sum contributions.

Proposition 7.5 (Finite many-to-one formula)**.** *Suppose X has density f_X and $Y = g(X)$ where for each relevant y the equation $g(x) = y$ has finitely many solutions $x_1, \dots, x_m$, each with $g'(x_i) \neq 0$. Then*

$$f_Y(y) = \sum_{i=1}^{m} \frac{f_X(x_i)}{|g'(x_i)|}.$$

Example 7.6 (Absolute value of a standard normal)**.** Let $X \sim \mathcal{N}(0,1)$ and $Y = |X|$. For $y > 0$, the equation $|x| = y$ has solutions $x = y$ and $x = -y$. Hence

$$f_Y(y) = \phi(y) + \phi(-y) = 2\phi(y), \qquad y > 0,$$

where ϕ is the standard normal density. This is the folded normal distribution.

Example 7.7 (Square of a standard normal)**.** If $X \sim \mathcal{N}(0,1)$ and $Y = X^2$, then for $y > 0$ the preimages are $\pm\sqrt{y}$, and

$$f_Y(y) = \frac{\phi(\sqrt{y}) + \phi(-\sqrt{y})}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \qquad y > 0.$$

This is the χ^2 distribution with one degree of freedom.

7.4 Sums of independent random variables and convolution

A particularly important transformation is the sum $S = X + Y$.

7.4.1 Discrete convolution

If X and Y are independent discrete random variables, then for any s,

$$\mathbb{P}(X + Y = s) = \sum_{x} \mathbb{P}(X = x)\mathbb{P}(Y = s - x).$$

This is called the *convolution* of the pmfs.

Example 7.8 (Sum of two fair dice)**.** If X and Y are independent uniform on $\{1,2,3,4,5,6\}$, then

$$\mathbb{P}(X + Y = 7) = \sum_{x} \mathbb{P}(X = x)\mathbb{P}(Y = 7 - x) = \frac{6}{36} = \frac{1}{6}.$$

The familiar triangular shape of the sum distribution comes from the number of decompositions of s as $x + (s - x)$.

7.4.2 Continuous convolution

If X and Y are independent continuous random variables with densities f_X and f_Y, then the density of $S = X + Y$ is

$$f_S(s) = \int_{-\infty}^{\infty} f_X(x) f_Y(s - x)\, dx.$$

Again, this is called convolution.

Example 7.9 (Sum of independent exponentials)**.** Let $X, Y \sim \text{Exp}(\lambda)$ be independent. Then for $s > 0$,

$$f_{X+Y}(s) = \int_0^s \lambda e^{-\lambda x} \lambda e^{-\lambda(s-x)}\, dx = \lambda^2 s e^{-\lambda s}.$$

So $X + Y$ has a gamma density with shape 2 and rate λ.

Convolution reflects a central principle in probability: distributions of sums are formed by averaging over all ways to split the sum. This idea will reappear in the study of generating functions, characteristic functions, and limit theorems.

7.5 Minima, maxima, and order statistics

Suppose $X_1, \ldots, X_n$ are i.i.d. with cdf F. The minimum and maximum often have especially simple distributions.

Proposition 7.10. *Let*

$$M_n = \max(X_1, \ldots, X_n), \qquad m_n = \min(X_1, \ldots, X_n).$$

Then

$$\mathbb{P}(M_n \le x) = F(x)^n,$$

and

$$\mathbb{P}(m_n > x) = (1 - F(x))^n, \qquad \textit{so} \qquad \mathbb{P}(m_n \le x) = 1 - (1 - F(x))^n.$$

Proof. For the maximum,

$$\{M_n \le x\} = \{X_1 \le x, \ldots, X_n \le x\}.$$

Because the variables are independent,

$$\mathbb{P}(M_n \le x) = \prod_{i=1}^{n} \mathbb{P}(X_i \le x) = F(x)^n.$$

The minimum is analogous. □

Example 7.11 (Maximum of uniforms)**.** If $X_1, \ldots, X_n$i.i.d.Unif$(0, 1)$, then for $0 < x < 1$,

$$\mathbb{P}(M_n \le x) = x^n,$$

so

$$f_{M_n}(x) = nx^{n-1}.$$

Likewise the minimum has density

$$f_{m_n}(x) = n(1-x)^{n-1}, \qquad 0 < x < 1.$$

These formulas are the simplest examples of *order statistics*. If we sort the sample into increasing order,

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)},$$

then $X_{(1)} = m_n$ and $X_{(n)} = M_n$. Order statistics matter in robust inference, simulation, and reliability theory.

7.5.1 The general density of an order statistic

For i.i.d. continuous random variables with density f and cdf F, the kth order statistic $X_{(k)}$ has density

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x).$$

The derivation is worth understanding heuristically. For $X_{(k)}$ to fall near x,

- one observation must land near x;
- exactly $k-1$ observations must lie below x;
- exactly $n-k$ must lie above x.

The combinatorial factor counts which observation occupies the kth position.

7.6 The Jacobian method in two dimensions

For transformations of pairs of continuous random variables, one uses a multidimensional change-of-variables formula.

Theorem 7.12 (Jacobian formula, informal version)**.** *Suppose (X, Y) has joint density $f_{X,Y}$ and define a one-to-one differentiable transformation*

$$(U, V) = g(X, Y)$$

with differentiable inverse $(X, Y) = h(U, V)$. Then

$$f_{U,V}(u, v) = f_{X,Y}(h(u, v))\,|\det J_h(u, v)|\,,$$

where J_h is the Jacobian matrix of the inverse map.

The determinant measures local area distortion. If small rectangles in the (u, v)-plane correspond to larger or smaller patches in the (x, y)-plane, that distortion must be accounted for.

Example 7.13 (Sum and difference)**.** Let

$$U = X + Y, \qquad V = X - Y.$$

Then the inverse transformation is

$$X = \frac{U + V}{2}, \qquad Y = \frac{U - V}{2},$$

whose Jacobian determinant has absolute value $1/2$. Therefore

$$f_{U,V}(u, v) = f_{X,Y}\left(\frac{u + v}{2}, \frac{u - v}{2}\right)\frac{1}{2}.$$

This transformation is especially useful when studying independent normal variables.

7.7 Simulation and the inverse-cdf method

Transformations are not only analytic tools; they are also computational tools. Suppose $U \sim \text{Unif}(0, 1)$ and F is a strictly increasing cdf. Then

$$X = F^{-1}(U)$$

has cdf F. Indeed,

$$\mathbb{P}(X \le x) = \mathbb{P}(F^{-1}(U) \le x) = \mathbb{P}(U \le F(x)) = F(x).$$

This is the *inverse-cdf method*. It reduces random generation from arbitrary distributions to random generation from the uniform distribution.

The method is especially helpful conceptually. It says that, in a precise sense, the uniform distribution on $(0, 1)$ is a universal raw source of randomness.

7.8 Summary

This chapter studied how distributions behave under deterministic transformations.

- The cdf method is the most general one-dimensional tool.
- Monotone differentiable transformations are handled by the change-of-variables formula.
- Many-to-one transformations require summing contributions from all preimages.
- Sums of independent variables are governed by convolution.

- Minima, maxima, and order statistics have clean formulas in the i.i.d. setting.
- In higher dimensions, the Jacobian determinant accounts for area distortion.

The next chapter returns to conditioning, but now at a deeper level: conditional expectation.

Exercises

Exercise 7.1. Let $X \sim \text{Unif}(0,1)$ and $Y = 1 - X$. Find the density of Y.

Exercise 7.2. Let $X \sim \text{Exp}(\lambda)$ and define $Y = 2X$. Find the density of Y.

Exercise 7.3. If $X \sim \mathcal{N}(0,1)$ and $Y = X^2$, derive the density of Y.

Exercise 7.4. Let X and Y be independent Bernoulli$(1/2)$ variables. Find the distribution of $X + Y$.

Exercise 7.5. Let X and Y be independent uniform random variables on $(0,1)$. Compute the density of $X + Y$.

Exercise 7.6. If $X_1, \ldots, X_n \text{i.i.d.} \text{Unif}(0,1)$, find the cdf and density of the minimum and maximum.

Exercise 7.7. Let $X_1, \ldots, X_n \text{i.i.d.}$ with cdf F. Show that the median order statistic in an odd sample has density given by the general order-statistic formula with $k = (n+1)/2$.

Exercise 7.8. Let (X, Y) be uniformly distributed on the unit square. Define $U = X + Y$ and $V = X - Y$. Find the inverse transformation and the absolute Jacobian determinant.

Challenge Exercise 7.9. Show that if $X, Y \text{i.i.d.} \text{Exp}(\lambda)$, then conditional on $X + Y = s$, the variable X is uniform on $(0, s)$.

Challenge Exercise 7.10. Suppose $X_1, \ldots, X_n \text{i.i.d.} \text{Exp}(\lambda)$. Show that the minimum m_n is exponential with rate $n\lambda$.

Chapter 8

Conditional Expectation

Conditional expectation is the numerical counterpart of conditional probability. Instead of asking for the chance of an event after information is revealed, we ask for the average value of a random quantity after information is revealed. It is one of the deepest and most useful ideas in probability.

8.1 From conditional probability to conditional averages

Conditional probability tells us how to update probabilities when we learn an event or a partial description of the outcome. But probabilities are only one kind of numerical question. We may also want to update an expected value.

For example, if X is the number of customers arriving today and Y is the number of customers arriving in the morning, then after learning Y we would like to know the best updated average for X. If Z is the total number of defective items in a batch and W is the number inspected in a first sample, then the average of Z should depend on the observed value of W. This is precisely what conditional expectation formalizes.

At the most basic level, if A is an event with positive probability and X is integrable, we define

$$\mathbb{E}[X \mid A] = \frac{\mathbb{E}[X\mathbf{1}_A]}{\mathbb{P}(A)}.$$

This is the average value of X restricted to the world in which A occurs.

Example 8.1. Roll a fair die and let X be the outcome. If A is the event that the result is even, then

$$\mathbb{E}[X \mid A] = \frac{2+4+6}{3} = 4.$$

Formally,

$$\mathbb{E}[X \mid A] = \frac{2\cdot(1/6) + 4\cdot(1/6) + 6\cdot(1/6)}{1/2} = 4.$$

This is already useful, but conditioning on a single event is too narrow for most applications. We usually condition on a random variable or on some body of information.

8.2 Conditioning on a discrete random variable

Suppose Y is a discrete random variable taking values $y_1, y_2, \dots$ with positive probability. For each possible value y, we can compute the conditional average of X given $Y = y$.

Definition 8.2. If X is integrable and Y is discrete, the *conditional expectation of X given $Y = y$* is

$$\mathbb{E}[X \mid Y = y] = \sum_x x\mathbb{P}(X = x \mid Y = y)$$

when X is discrete, or the corresponding integral if X is continuous conditional on $Y = y$.

This quantity depends on y, so it defines a function of the observed value. We therefore package the whole family into a new random variable.

Definition 8.3. If Y is discrete, the *conditional expectation of X given Y* is the random variable

$$\mathbb{E}[X \mid Y] = g(Y), \qquad \text{where } g(y) = \mathbb{E}[X \mid Y = y].$$

Thus $\mathbb{E}[X \mid Y]$ is itself a random variable. Before observing Y, it is random; after observing $Y = y$, it becomes the number $g(y)$.

Example 8.4 (Binomial given total trials)**.** Let N be a Poisson random variable representing the number of opportunities, and suppose that conditional on $N = n$, each opportunity independently succeeds with probability p. Let S be the total number of successes. Then conditional on $N = n$, we have $S \sim \text{Bin}(n, p)$, so

$$\mathbb{E}[S \mid N = n] = np.$$

Hence

$$\mathbb{E}[S \mid N] = pN.$$

This is an excellent example of the "replace the random parameter by the observed value" rule that often appears in conditional expectation calculations.

8.3 The defining properties of conditional expectation

The notation $\mathbb{E}[X \mid Y]$ suggests an average depending on Y, but what properties characterize it? At undergraduate level, the right answer can be stated quite concretely.

Theorem 8.5 (Characterization given a random variable)**.** *Let X be integrable and Y any random variable. A random variable Z is a version of $\mathbb{E}[X \mid Y]$ if and only if:*

(i) Z is a function of Y;

(ii) for every reasonable function h of Y for which the expectations exist,

$$\mathbb{E}[h(Y)Z] = \mathbb{E}[h(Y)X].$$

The first property says that Z uses only the information contained in Y. The second says that, from the perspective of all events determined by Y, the random variable Z behaves exactly like X on average. In more advanced language, property (ii) is the defining integral identity.

We will not prove the full existence theorem here. In discrete settings one can build the conditional expectation directly. In general settings it is a theorem from measure theory. What matters for us is how to use the object once we have it.

8.4 Fundamental properties

Conditional expectation has a small set of structural rules that account for most computations.

Theorem 8.6 (Basic properties)**.** *Assume the expectations below exist.*

(1) ***Linearity:***

$$\mathbb{E}[aX + bW \mid Y] = a\mathbb{E}[X \mid Y] + b\mathbb{E}[W \mid Y].$$

(2) ***Taking out what is known:*** *if $g(Y)$ is a function of Y, then*

$$\mathbb{E}[g(Y)X \mid Y] = g(Y)\mathbb{E}[X \mid Y].$$

(3) ***Positivity:*** *if $X \geq 0$ almost surely, then $\mathbb{E}[X \mid Y] \geq 0$ almost surely.*

(4) ***Tower property:***

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

(5) ***Independence:*** *if X is independent of Y, then*

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X].$$

Remark 8.7. These properties look natural, but they are powerful enough to drive a large part of modern probability. The tower property alone appears in essentially every serious stochastic argument.

Why the tower property is plausible. If Y partitions the world into cases, then $\mathbb{E}[X \mid Y]$ is the average of X within each case. Averaging those case-by-case averages over the distribution of Y should return the global average. In the discrete case,

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) = \mathbb{E}[X],$$

which is exactly the law of total expectation. □

8.5 Conditioning on a partition

A very useful special case is conditioning on a finite or countable partition.

Suppose $B_1, B_2, \dots$ is a partition of Ω with positive probabilities. Define the random variable

$$Z = \sum_i \mathbb{E}[X \mid B_i]\mathbf{1}_{B_i}.$$

Then Z is the conditional expectation of X with respect to the information "which block of the partition occurred." This is nothing more than the random-variable version of averaging over cases.

Example 8.8 (Two-component mixture)**.** A store has a low-traffic day with probability 0.7 and a high-traffic day with probability 0.3. Let N be the number of customers. On low-traffic days, $\mathbb{E}[N] = 20$; on high-traffic days, $\mathbb{E}[N] = 50$. If D indicates the day type, then

$$\mathbb{E}[N \mid D] = 20\mathbf{1}_{\{D=L\}} + 50\mathbf{1}_{\{D=H\}}.$$

Taking expectations,

$$\mathbb{E}[N] = 0.7 \cdot 20 + 0.3 \cdot 50 = 29.$$

8.6 Conditional expectation and best prediction

One of the deepest ways to understand conditional expectation is as an optimal predictor.

Suppose we observe Y and must predict X using only information from Y. Among all predictors of the form $g(Y)$, which one minimizes mean squared error

$$\mathbb{E}[(X - g(Y))^2]?$$

The answer is conditional expectation.

Theorem 8.9 (Least-squares characterization)**.** *If X has finite second moment, then among all square-integrable functions $g(Y)$,*

$$\mathbb{E}[(X - g(Y))^2]$$

is minimized when $g(Y) = \mathbb{E}[X \mid Y]$.

Proof sketch. Write

$$X - g(Y) = (X - \mathbb{E}[X \mid Y]) + (\mathbb{E}[X \mid Y] - g(Y)).$$

Square and expand. The cross term has expectation 0 because $\mathbb{E}[X \mid Y] - g(Y)$ is a function of Y, while $X - \mathbb{E}[X \mid Y]$ has conditional mean zero given Y. Thus

$$\mathbb{E}[(X - g(Y))^2] = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] + \mathbb{E}[(\mathbb{E}[X \mid Y] - g(Y))^2].$$

The second term is nonnegative and is zero exactly when $g(Y) = \mathbb{E}[X \mid Y]$ almost surely. □

This identity is sometimes called the *orthogonal decomposition* of prediction error. It has a geometric flavor: conditional expectation is the projection of X onto the space of functions of Y.

8.7 Conditional variance and the variance decomposition formula

Conditional expectation leads naturally to conditional variance.

Definition 8.10. The *conditional variance* of X given Y is

$$\operatorname{Var}(X \mid Y) = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2 \mid Y\big].$$

The fundamental identity is the variance decomposition formula.

Theorem 8.11 (Law of total variance)**.** *If X has finite second moment, then*

$$\operatorname{Var}(X) = \mathbb{E}[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(\mathbb{E}[X \mid Y]).$$

Proof. Add and subtract $\mathbb{E}[X \mid Y]$ inside $X - \mathbb{E}[X]$:

$$X - \mathbb{E}[X] = (X - \mathbb{E}[X \mid Y]) + (\mathbb{E}[X \mid Y] - \mathbb{E}[X]).$$

Square both sides and take expectations. The mixed term vanishes by the same reasoning as in the least-squares proof. The two remaining terms are exactly the terms in the formula. □

This theorem separates total variability into two pieces:

- variability left over even after Y is known;
- variability in the conditional mean itself across different values of Y.

It is one of the conceptual foundations of regression, ANOVA, and hierarchical modeling.

8.8 Examples of calculation

8.8.1 Conditioning on a binomial count

Suppose $N \sim \operatorname{Bin}(n, p)$ and, conditional on $N = k$, a variable X has mean $2k + 1$. Then

$$\mathbb{E}[X \mid N] = 2N + 1,$$

and therefore

$$\mathbb{E}[X] = 2\mathbb{E}[N] + 1 = 2np + 1.$$

The lesson is simple but powerful: if the conditional expectation is easy to describe, then the unconditional expectation follows immediately by the tower property.

8.8.2 Competing exponentials

Let $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\mu)$ be independent. Define $T = \min(X, Y)$ and let $I = \mathbf{1}_{\{X<Y\}}$. Then

$$\mathbb{P}(I = 1) = \mathbb{P}(X < Y) = \frac{\lambda}{\lambda + \mu}.$$

One way to see this is by conditioning on X or by integrating the joint density over the region $x < y$. Here conditional expectation packages the argument elegantly: I is an indicator, so

$$\mathbb{E}[I] = \mathbb{P}(X < Y).$$

Conditioning on X gives

$$\mathbb{E}[I \mid X] = \mathbb{P}(Y > X \mid X) = e^{-\mu X}.$$

Hence

$$\mathbb{P}(X < Y) = \mathbb{E}[e^{-\mu X}] = \int_0^\infty e^{-\mu x} \lambda e^{-\lambda x}\, dx = \frac{\lambda}{\lambda + \mu}.$$

This is a good example of conditional expectation turning a probability question into an expectation question and then into a manageable integral.

8.8.3 Random sums

Suppose $X_1, X_2, \ldots$ are i.i.d. with mean m, and let N be an independent nonnegative integer-valued random variable. Consider the random sum

$$S_N = X_1 + \cdots + X_N,$$

with the convention $S_0 = 0$. Then

$$\mathbb{E}[S_N \mid N] = Nm,$$

and therefore

$$\mathbb{E}[S_N] = m\mathbb{E}[N].$$

This identity, known as Wald's equation in a simple form, appears throughout applied probability.

8.9 Conditional expectation given a σ-field

For conceptual completeness, we briefly state the most general version.

A collection of information is often represented not by a random variable but by a σ-field $\mathcal{G}$. The conditional expectation of X given $\mathcal{G}$, written $\mathbb{E}[X \mid \mathcal{G}]$, is the random variable that is measurable with respect to $\mathcal{G}$ and satisfies

$$\mathbb{E}[\mathbf{1}_A \mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[\mathbf{1}_A X] \qquad \text{for all } A \in \mathcal{G}.$$

You should think of $\mathcal{G}$ as "the information currently available." If Y is a random variable, then

conditioning on Y is the same as conditioning on the information generated by Y, usually denoted $\sigma(Y)$. Thus

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \sigma(Y)].$$

At this stage we will not need the full formal machinery often, but this language is worth seeing because it unifies many special cases and prepares the ground for later study of stochastic processes.

8.10 Jensen's inequality for conditional expectation

One more property deserves mention because it is conceptually important.

Theorem 8.12 (Conditional Jensen inequality)**.** *If φ is convex and X is integrable with $\varphi(X)$ integrable, then*

$$\varphi(\mathbb{E}[X \mid Y]) \leq \mathbb{E}[\varphi(X) \mid Y]$$

almost surely.

This is the conditional analog of the usual Jensen inequality. It says that averaging before applying a convex function cannot increase the result. Among other things, it implies

$$(\mathbb{E}[X \mid Y])^2 \leq \mathbb{E}[X^2 \mid Y].$$

This inequality is often useful in moment bounds.

8.11 How to compute conditional expectations in practice

In actual problem solving, the following strategies are the most common.

(1) **Condition on the most informative simple variable.** If a random parameter drives the rest of the model, condition on that parameter first.

(2) **Use indicators.** To compute conditional probabilities, write them as conditional expectations of indicators.

(3) **Exploit symmetry.** If conditional on one variable all remaining possibilities are symmetric, the conditional mean often has a very simple form.

(4) **Guess the functional form.** If $\mathbb{E}[X \mid Y]$ must be a function of Y, sometimes it is easy to guess which one and verify it using the defining property.

(5) **Remember that independence collapses conditioning.**

8.12 Summary

Conditional expectation is one of the structural pillars of probability.

- It generalizes conditional probability from events to random quantities.
- It is itself a random variable, typically a function of the observed information.
- The tower property, linearity, and "taking out what is known" are its most important rules.
- It is the best mean-square predictor based on the available information.
- The law of total variance decomposes variability into explained and unexplained parts.

The next chapter turns to generating functions and characteristic functions, which provide algebraic tools for understanding distributions and sums of independent random variables.

Exercises

Exercise 8.1. A fair die is rolled. Let X be the outcome and let A be the event that the outcome is at least 4. Compute $\mathbb{E}[X \mid A]$.

Exercise 8.2. Let X and Y be jointly discrete with

$$\mathbb{P}(X = x, Y = y) = \frac{1}{8}, \qquad (x, y) \in \{0, 1\} \times \{0, 1, 2, 3\},$$

arranged so that Y is uniform on $\{0, 1, 2, 3\}$ and $X = 1$ exactly when $Y \in \{2, 3\}$. Compute $\mathbb{E}[X \mid Y]$.

Exercise 8.3. Suppose $N \sim \text{Pois}(\lambda)$ and, conditional on $N = n$, the variable X is binomial $\text{Bin}(n, p)$. Compute $\mathbb{E}[X \mid N]$ and $\mathbb{E}[X]$.

Exercise 8.4. Let X and Y be independent with $\mathbb{E}[X] = 5$. Show that $\mathbb{E}[X \mid Y] = 5$.

Exercise 8.5. Let X be integrable and Y discrete. Prove directly that

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

Exercise 8.6. Suppose X has finite second moment and Y is any random variable. Prove that

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]).$$

Exercise 8.7. Let $X_1, X_2, \ldots$ be i.i.d. with mean m, and let N be independent of the sequence with finite mean. Use conditioning on N to prove that

$$\mathbb{E}[X_1 + \cdots + X_N] = m\mathbb{E}[N].$$

Exercise 8.8. If X is nonnegative and A is an event with positive probability, show that

$$\mathbb{E}[X \mid A] \geq 0.$$

Give an example in which the inequality is strict even though $\mathbb{E}[X] = 0$.

Challenge Exercise 8.9. Let $X \sim \text{Exp}(1)$ and define $Y = \lfloor X \rfloor$, the integer part of X. Compute $\mathbb{E}[X \mid Y]$.

Challenge Exercise 8.10. Suppose X and Y are square-integrable. Show that among all constants c, the value of c minimizing $\mathbb{E}[(X - c)^2]$ is $c = \mathbb{E}[X]$. Then explain how this is the simplest case of the least-squares characterization of conditional expectation.

Chapter 9

Generating Functions and Characteristic Functions

Generating functions encode distributions into ordinary functions. This translation turns probabilistic operations such as summation, convolution, and moment calculation into algebra and calculus. For discrete counts, probability generating functions are especially natural; for general random variables, characteristic functions provide a universal Fourier-analytic tool.

9.1 Why generating functions are useful

A distribution can be described by a pmf, a density, or a cdf. Generating functions provide a fourth description, one that often makes algebraic structure more transparent. The basic philosophy is simple: instead of storing probabilities directly, package them into a function whose coefficients or derivatives recover the probabilistic quantities of interest.

This is useful for at least three reasons.

(1) Products of generating functions correspond to sums of independent random variables.

(2) Derivatives recover moments.

(3) Analytic properties of the generating function often reflect probabilistic properties of the distribution.

The right generating function depends on the setting. For nonnegative integer-valued variables, the probability generating function is the cleanest. For variables with exponential moments, moment generating functions are very convenient. For all real-valued variables, characteristic functions always exist and are therefore the most robust.

9.2 Probability generating functions

Definition 9.1. If X is a nonnegative integer-valued random variable, its *probability generating function* (pgf) is

$$G_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} \mathbb{P}(X = k)s^k, \qquad |s| \le 1.$$

The pgf is literally an ordinary power series whose coefficients are the probabilities $\mathbb{P}(X = k)$. Because the coefficients sum to 1, the series converges absolutely for $|s| \le 1$.

Some immediate facts are worth recording.

- $G_X(1) = 1$.
- $G_X(0) = \mathbb{P}(X = 0)$.
- If the derivatives exist at 1, then they encode moments and factorial moments.

Proposition 9.2. *If X is nonnegative integer-valued and $\mathbb{E}[X] < \infty$, then*

$$G'_X(1) = \mathbb{E}[X].$$

If $\mathbb{E}[X(X-1)] < \infty$, then

$$G''_X(1) = \mathbb{E}[X(X-1)].$$

Hence

$$\operatorname{Var}(X) = G''_X(1) + G'_X(1) - \left(G'_X(1)\right)^2.$$

Proof. Differentiate term by term:

$$G'_X(s) = \sum_{k=1}^{\infty} k\mathbb{P}(X = k)s^{k-1}.$$

Setting $s = 1$ gives the first identity. The second is similar. □

Example 9.3 (Binomial pgf)**.** If $X \sim \mathrm{Bin}(n, p)$, then

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

The binomial theorem makes the pgf calculation essentially automatic.

Example 9.4 (Poisson pgf)**.** If $X \sim \mathrm{Pois}(\lambda)$, then

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

9.3 Sums of independent integer-valued variables

Generating functions turn convolution into multiplication.

Theorem 9.5. *If X and Y are independent nonnegative integer-valued random variables, then*

$$G_{X+Y}(s) = G_X(s)G_Y(s).$$

More generally, for independent $X_1, \dots, X_n$,

$$G_{X_1+\cdots+X_n}(s) = \prod_{i=1}^{n} G_{X_i}(s).$$

Proof. By independence,

$$G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X s^Y] = \mathbb{E}[s^X]\mathbb{E}[s^Y] = G_X(s)G_Y(s).$$

□

This identity explains many closure properties of classical families.

Example 9.6 (Poisson additivity)**.** If $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ are independent, then

$$G_{X+Y}(s) = e^{\lambda(s-1)}e^{\mu(s-1)} = e^{(\lambda+\mu)(s-1)},$$

so

$$X + Y \sim \text{Pois}(\lambda + \mu).$$

Example 9.7 (Branching-process preview)**.** Suppose each individual in a population has a random number of offspring with pgf $G(s)$, independently of the others. If the current generation has size n, then the total size of the next generation has pgf $G(s)^n$. This simple observation drives the theory of Galton–Watson branching processes.

9.4 Moment generating functions

The pgf is tied to integer-valued nonnegative variables. For general real-valued variables, a natural alternative is the moment generating function.

Definition 9.8. The *moment generating function* (mgf) of a random variable X is

$$M_X(t) = \mathbb{E}[e^{tX}],$$

for those real numbers t for which the expectation exists.

The word "moment" comes from the fact that derivatives at 0 recover moments, at least when differentiation under the expectation is justified:

$$M_X'(0) = \mathbb{E}[X], \qquad M_X''(0) = \mathbb{E}[X^2], \qquad \text{etc.}$$

Example 9.9. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

This can be shown by completing the square in the integral.

Example 9.10. If $X \sim \text{Exp}(\lambda)$, then for $t < \lambda$,

$$M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \frac{\lambda}{\lambda - t}.$$

The mgf does not exist for $t \geq \lambda$, which is a reminder that mgfs need not exist globally.

Theorem 9.11. *If X and Y are independent and the mgfs exist in a neighborhood of 0, then*

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Thus mgfs provide the same multiplicative simplification for sums of independent variables that pgfs do in the counting setting.

9.5 Uniqueness and limitations of mgfs

If an mgf exists on an open interval around 0, it uniquely determines the distribution. This is a major advantage. However, not every random variable has an mgf in such a neighborhood. Heavy-tailed distributions often fail this requirement.

Example 9.12 (Cauchy distribution)**.** The Cauchy distribution has no finite mean and no mgf near 0. Thus mgfs, while powerful, are not universal.

Because of this limitation, probability theory uses a more robust transform: the characteristic function.

9.6 Characteristic functions

Definition 9.13. The *characteristic function* of a real-valued random variable X is

$$\varphi_X(t) = \mathbb{E}[e^{itX}], \qquad t \in \mathbb{R},$$

where $i = \sqrt{-1}$.

This is simply the Fourier transform of the distribution. Since

$$|e^{itX}| = 1,$$

the expectation always exists. That is the crucial advantage: every real-valued random variable has a characteristic function.

Example 9.14. If $X \sim \mathrm{Ber}(p)$, then

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = (1-p) + pe^{it}.$$

If $X \sim \mathrm{Pois}(\lambda)$, then

$$\varphi_X(t) = e^{\lambda(e^{it}-1)}.$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\varphi_X(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right).$$

9.6.1 Basic properties

Characteristic functions satisfy several simple but important rules.

(1) $\varphi_X(0) = 1$.

(2) $|\varphi_X(t)| \le 1$ for all t.

(3) If $a, b \in \mathbb{R}$, then

$$\varphi_{aX+b}(t) = e^{itb}\varphi_X(at).$$

(4) If X and Y are independent, then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

The last property is the key reason characteristic functions are useful in limit theorems.

9.7 Moments from characteristic functions

If X has enough moments, derivatives of φ_X at 0 recover them. For example, when $\mathbb{E}|X| < \infty$,

$$\varphi_X'(0) = i\mathbb{E}[X].$$

If $\mathbb{E}[X^2] < \infty$, then

$$\varphi_X''(0) = -\mathbb{E}[X^2].$$

Thus the characteristic function not only identifies the distribution but also packages moment information.

9.8 Uniqueness and convergence

Two fundamental theorems make characteristic functions central in probability.

Theorem 9.15 (Uniqueness)**.** *If two random variables have the same characteristic function, then they have the same distribution.*

Theorem 9.16 (Continuity theorem, informal version)**.** *If $\varphi_{X_n}(t) \to \varphi(t)$ for every t, and if φ is continuous at 0, then there exists a random variable X with characteristic function φ, and*

$$X_n \xrightarrow{d} X.$$

The first theorem says characteristic functions encode all distributional information. The second says convergence of characteristic functions implies convergence in distribution. Together, they make characteristic functions ideal for proving the central limit theorem.

A full proof of the continuity theorem requires more Fourier analysis than we want in this course. Still, it is important to know the statement and to become comfortable using it.

9.9 Applications to sums and approximations

9.9.1 A quick proof of Poisson additivity

If $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ are independent, then

$$\varphi_{X+Y}(t) = e^{\lambda(e^{it}-1)} e^{\mu(e^{it}-1)} = e^{(\lambda+\mu)(e^{it}-1)}.$$

By uniqueness, $X + Y \sim \text{Pois}(\lambda + \mu)$.

9.9.2 Normal sums

If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then

$$\varphi_{X+Y}(t) = e^{it\mu_1 - \sigma_1^2 t^2/2} e^{it\mu_2 - \sigma_2^2 t^2/2} = e^{it(\mu_1+\mu_2) - (\sigma_1^2+\sigma_2^2)t^2/2}.$$

Hence

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

This is one of the defining stability properties of the normal family.

9.10 Characteristic functions and the central limit heuristic

Suppose $X_1, X_2, \dots$ are i.i.d. with mean 0 and variance 1, and define

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

Then, by independence,

$$\varphi_{S_n}(t) = \left(\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

If $\varphi_X(u) = 1 - u^2/2 + o(u^2)$ near 0, then

$$\varphi_{S_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-t^2/2}.$$

But $e^{-t^2/2}$ is the characteristic function of $\mathcal{N}(0,1)$. This is the basic engine behind the central limit theorem. The full theorem will be proved in the next major asymptotic chapter, but the structure is already visible here.

9.11 A note on style: transforms are tools, not ends in themselves

Students sometimes find generating functions mysterious because they seem to replace a concrete distribution with a more abstract object. The right perspective is practical: transforms are not meant to obscure the distribution but to make hidden structure visible. If a problem is easy directly from the pmf or density, do it directly. Use a generating function when it simplifies sums, products, or asymptotics.

This is also why several different transforms coexist. Each is adapted to a different family of problems.

- **pgf:** best for counts and branching structures.
- **mgf:** best for moment calculations and some exponential tail arguments.
- **characteristic function:** best for universality and limit theorems.

9.12 Summary

This chapter introduced analytic encodings of distributions.

- The pgf packages probabilities of integer-valued variables into a power series.
- The mgf packages moments via exponential tilting, when it exists.
- The characteristic function always exists and uniquely determines the distribution.

- Products of transforms correspond to sums of independent random variables.
- Characteristic functions provide the main bridge to the central limit theorem.

The next chapter begins the asymptotic part of the course by studying modes of convergence and the laws of large numbers.

Exercises

Exercise 9.1. Compute the pgf of a geometric random variable with parameter p in the convention $\mathbb{P}(X = k) = (1-p)^{k-1}p$, $k \geq 1$.

Exercise 9.2. Use the pgf of a Poisson random variable to compute its mean and variance.

Exercise 9.3. Show directly from mgfs that the sum of independent normal random variables is normal.

Exercise 9.4. Compute the mgf of a Bernoulli(p) random variable.

Exercise 9.5. Let X and Y be independent nonnegative integer-valued random variables. Prove that the pgf of $X + Y$ is the product of the pgfs.

Exercise 9.6. Find the characteristic function of a uniform distribution on $(-1, 1)$.

Exercise 9.7. If X is symmetric about 0, show that its characteristic function is real-valued and even.

Exercise 9.8. Suppose $X \sim \text{Pois}(\lambda)$ and, conditional on X, each of the X points is kept independently with probability p. Let Y be the number kept. Use pgfs to show that $Y \sim \text{Pois}(\lambda p)$.

Challenge Exercise 9.9. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) and let S_n be their sum. Use characteristic functions to identify the distribution of S_n without referring to direct counting.

Challenge Exercise 9.10. Suppose X takes values in $\{0, 1, 2, \ldots\}$ and has pgf $G(s) = \exp\{\lambda(s-1)\}$. Show that X must be Poisson with parameter λ.

Chapter 10

Modes of Convergence and the Laws of Large Numbers

Probability theory studies not only single random variables but also sequences of random variables. The laws of large numbers describe how averages stabilize as sample size grows. To state them precisely, we need several notions of convergence: almost sure, in probability, in mean square, and in distribution.

10.1 Why convergence matters

A single random experiment may be noisy, but repeated averaging often reveals stable structure. This empirical fact lies behind much of statistics. Sample proportions stabilize. Sample means become reliable summaries. Simulations converge. Frequencies begin to resemble probabilities.

The laws of large numbers turn these observations into mathematics. But there is more than one way for a sequence of random variables to converge. Each mode of convergence captures a different kind of asymptotic behavior, and learning to distinguish them is essential.

10.2 Almost sure convergence

Definition 10.1. A sequence of random variables X_n converges *almost surely* to a random variable X, written

$$X_n \xrightarrow{a.s.} X,$$

if

$$\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1.$$

This is the strongest common mode of probabilistic convergence. It means that for all outcomes except possibly those in a probability-zero exceptional set, the numerical sequence $X_n(\omega)$ converges ordinarily.

A useful way to think about almost sure convergence is pathwise convergence. Fix an outcome ω and watch the sample path. If the path converges for almost every ω, then we have almost sure convergence.

Example 10.2. Let $X_n(\omega) = \omega^n$ on $\Omega = (0,1)$ with the uniform distribution. Then for every $\omega \in (0,1)$, $\omega^n \to 0$. Hence $X_n \xrightarrow{a.s.} 0$.

10.3 Convergence in probability

Definition 10.3. A sequence X_n converges *in probability* to X, written

$$X_n \xrightarrow{\mathbb{P}} X,$$

if for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0.$$

This says that large deviations from the target become unlikely. It does not require pointwise convergence on individual sample paths.

Example 10.4. Suppose A_n are events with $\mathbb{P}(A_n) = 1/n$, and define $X_n = \mathbf{1}_{A_n}$. Then for every $\varepsilon \in (0,1)$,

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n = 1) = \frac{1}{n} \to 0.$$

So $X_n \xrightarrow{\mathbb{P}} 0$. Whether $X_n \xrightarrow{a.s.} 0$ depends on the structure of the events A_n; convergence in probability alone does not decide it.

10.4 Convergence in distribution

Definition 10.5. A sequence X_n converges *in distribution* to X, written

$$X_n \xrightarrow{d} X,$$

if

$$F_{X_n}(x) \to F_X(x)$$

at every continuity point x of F_X.

Convergence in distribution is weaker than convergence in probability. It concerns only the limiting shape of the distributions, not whether the variables can be coupled on the same sample space to be numerically close.

Example 10.6. If $X_n \sim \mathcal{N}(0, 1 + 1/n)$, then $X_n \xrightarrow{d} Z$ where $Z \sim \mathcal{N}(0,1)$. Even if the variables are defined on different probability spaces, the distributional statement still makes sense.

10.5 Implication structure

The common modes of convergence are related as follows.

Theorem 10.7. *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$. If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{d} X$.*

The converses are false in general. These failures are important and instructive.

Proof that almost sure convergence implies convergence in probability. Fix $\varepsilon > 0$. If $X_n \to X$ almost surely, then the indicators

$$\mathbf{1}_{\{|X_n - X| > \varepsilon\}}$$

converge almost surely to 0. Since these indicators are bounded by 1, the dominated convergence theorem implies

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}\left[\mathbf{1}_{\{|X_n - X| > \varepsilon\}}\right] \to 0.$$

Thus $X_n \xrightarrow{\mathbb{P}} X$. □

Remark 10.8. Convergence in distribution to a constant is actually equivalent to convergence in probability to that constant. This small but important theorem is often used in limit arguments.

10.6 Convergence in mean square and L^p

Sometimes one controls moments of the difference directly.

Definition 10.9. We say that X_n converges to X in mean square, or in L^2, if

$$\mathbb{E}[(X_n - X)^2] \to 0.$$

More generally, convergence in L^p means

$$\mathbb{E}[|X_n - X|^p] \to 0.$$

By Markov's inequality, convergence in L^p implies convergence in probability. In particular, mean-square convergence is strong enough to imply convergence in probability.

10.7 The weak law of large numbers

We now turn to averages. Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ. The sample average is

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

The most basic stabilization statement is the weak law.

Theorem 10.10 (Weak law of large numbers, finite variance version)**.** *Suppose $X_1, X_2, \dots$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$. Then*

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Proof. By linearity of expectation,

$$\mathbb{E}[\overline{X}_n] = \mu.$$

Because the variables are independent,

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

Now apply Chebyshev's inequality:

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \le \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

□

This proof is short but profound. It shows how variance control plus Chebyshev's inequality yields asymptotic concentration. It is the prototype of many later probabilistic arguments.

Remark 10.11. The weak law remains true under weaker assumptions than finite variance. For instance, i.i.d. integrable random variables satisfy a weak law. Proving that more general result requires ideas beyond this elementary argument.

10.8 Interpretation of the weak law

The weak law does not say that every sample path average converges. It says that for large n, the sample average is very likely to be close to the true mean. This is exactly the level of statement needed for many statistical applications.

As an example, if $X_i \sim \mathrm{Ber}(p)$, then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the sample proportion of successes. The weak law says

$$\overline{X}_n \xrightarrow{\mathbb{P}} p.$$

So observed frequencies become good approximations to the underlying probability parameter.

10.9 Borel–Cantelli lemmas

To reach almost sure convergence, we need a tool that controls the occurrence of infinitely many rare events.

Theorem 10.12 (First Borel–Cantelli lemma)**.** *If $A_1, A_2, \dots$ are events and*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then

$$\mathbb{P}(A_n \text{ infinitely often}) = 0.$$

Here "A_n infinitely often" means that infinitely many of the events occur.

Proof. Let

$$B_m = \bigcup_{n \geq m} A_n.$$

Then the events B_m decrease to the event that infinitely many A_n occur. By the union bound,

$$\mathbb{P}(B_m) \leq \sum_{n \geq m} \mathbb{P}(A_n).$$

Since the tail of the convergent series tends to 0, we conclude that $\mathbb{P}(B_m) \to 0$, and hence the limit event has probability 0. □

The converse direction requires independence.

Theorem 10.13 (Second Borel–Cantelli lemma)**.** *If $A_1, A_2, \dots$ are independent and*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$

then

$$\mathbb{P}(A_n \text{ infinitely often}) = 1.$$

We will not prove the second lemma in full detail here, but its message is intuitive: independent events whose probabilities do not sum to a finite number must keep occurring forever.

10.10 The strong law of large numbers

Almost sure convergence of sample averages is the strong law.

Theorem 10.14 (Strong law of large numbers, finite variance version)**.** *Suppose $X_1, X_2, \dots$ are i.i.d. with mean μ and finite variance. Then*

$$\overline{X}_n \xrightarrow{a.s.} \mu.$$

The full proof is more involved than the weak law. There are several routes, each teaching something useful. We present a classical strategy in outline.

10.10.1 Idea of the proof

Set $S_n = X_1 + \cdots + X_n$ and first assume $\mu = 0$ for simplicity. The goal is to show $S_n/n \to 0$ almost surely.

One efficient route is:

(1) show that along the dyadic subsequence $n = 2^k$, the quantities $S_{2^k}/2^k$ converge almost surely to 0;

(2) use maximal inequalities and control of increments to pass from the subsequence to all n.

The dyadic subsequence is easier because

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{2^k}}{2^k}\right| > \varepsilon\right) \le \sum_{k=1}^{\infty} \frac{\operatorname{Var}(S_{2^k})}{\varepsilon^2 2^{2k}} = \sum_{k=1}^{\infty} \frac{2^k \sigma^2}{\varepsilon^2 2^{2k}} = \frac{\sigma^2}{\varepsilon^2} \sum_{k=1}^{\infty} 2^{-k} < \infty.$$

By the first Borel–Cantelli lemma,

$$\frac{S_{2^k}}{2^k} \to 0 \quad \text{almost surely.}$$

The second step requires showing that for $2^k \le n < 2^{k+1}$, the difference between S_n/n and $S_{2^k}/2^k$ is negligible. This can be done using bounds on block increments. The details are standard but a bit technical, so we omit them here.

Remark 10.15. The strongest classical version of the strong law requires only $\mathbb{E}|X_1| < \infty$. The finite-variance version above is easier to motivate and already captures the key phenomenon: repeated averaging stabilizes almost surely.

10.11 Sample means and empirical frequencies

The laws of large numbers justify the statistical practice of estimating a mean by a sample average.

- If $X_i \sim \text{Ber}(p)$, then the sample proportion estimates p.
- If X_i are measurements with mean μ, then the sample mean estimates μ.
- If X_i are costs, returns, waiting times, or defects, then their average estimates the long-run average performance of the system.

The weak law says the estimate is likely to be good for large n; the strong law says that along almost every infinite run of the experiment, the estimate eventually settles near the truth.

10.12 Rates and inequalities

The laws of large numbers are qualitative statements. They say convergence happens, but not how fast. The Chebyshev proof gives at least a crude rate in probability:

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

This is often enough to see how large a sample size must be to make a deviation event unlikely.

Sharper bounds are possible under stronger assumptions. For bounded independent variables, Hoeffding-type inequalities give exponentially small tails. Since modern data science frequently uses such concentration bounds, it is worth recording one representative statement.

Theorem 10.16 (Hoeffding inequality, optional)**.** *If $X_1, \dots, X_n$ are independent with $a_i \leq X_i \leq b_i$ almost surely, then for all $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

We will not prove this theorem here, but it is useful to know that much stronger concentration is available when boundedness or sub-Gaussian assumptions hold.

10.13 Convergence of empirical means versus convergence of distributions

It is important not to confuse two different asymptotic statements.

- The laws of large numbers concern convergence of the sample average to a constant.
- The central limit theorem concerns the distributional fluctuations of the centered and scaled sample average.

The law of large numbers tells us the average settles down. The central limit theorem, coming next, tells us what the residual fluctuations look like before they disappear under normalization.

10.14 Useful examples and counterexamples

10.14.1 A sequence converging in probability but not almost surely

Let $U \sim \mathrm{Unif}(0,1)$ and define

$$X_n = \mathbf{1}_{I_n}(U),$$

where the intervals I_n slide through $(0,1)$ in such a way that each has length going to 0 but every point lies in infinitely many of them. Then $\mathbb{P}(X_n = 1) = |I_n| \to 0$, so $X_n \xrightarrow{\mathbb{P}} 0$, yet X_n does not converge almost surely. This shows that convergence in probability need not be pathwise.

10.14.2 A sequence converging in distribution but not in probability

If X_n are i.i.d. standard normal, then each X_n has the same distribution as a standard normal Z, so $X_n \xrightarrow{d} Z$. But unless the variables are specially coupled, they do not converge in probability to Z or to anything else. The distributions stabilize while the actual values do not.

These examples clarify why the hierarchy of convergence modes is strict.

10.15 Summary

This chapter introduced the language of random-variable convergence and established the laws of large numbers.

- Almost sure convergence is pathwise convergence except on a null set.
- Convergence in probability means that large errors become unlikely.
- Convergence in distribution concerns only limiting distributions.
- The weak law states that sample averages converge in probability to the mean.
- The strong law upgrades this to almost sure convergence under suitable moment assumptions.
- Borel–Cantelli lemmas help control infinitely many rare events.

The next chapter studies the fluctuations around the law of large numbers: the central limit theorem.

Exercises

Exercise 10.1. Let $X_n = 1/n$. Show that $X_n \xrightarrow{a.s.} 0$, $X_n \xrightarrow{\mathbb{P}} 0$, and $X_n \xrightarrow{d} 0$.

Exercise 10.2. Let $X_n = \mathbf{1}_{A_n}$ where $\mathbb{P}(A_n) = 1/n^2$. Show that $X_n \xrightarrow{\mathbb{P}} 0$. Use the first Borel–Cantelli lemma to show that $X_n \xrightarrow{a.s.} 0$.

Exercise 10.3. Suppose $X_n \xrightarrow{\mathbb{P}} X$ and $Y_n \xrightarrow{\mathbb{P}} Y$. Show that $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$.

Exercise 10.4. Let $X_1, X_2, \dots$ be i.i.d. Bernoulli(p) variables. Use Chebyshev's inequality to prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} p.$$

Exercise 10.5. Suppose $X_1, \dots, X_n$ are independent with mean 0 and variance 1. Show that

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}.$$

Then derive a Chebyshev bound.

Exercise 10.6. State carefully the difference between convergence in probability and convergence in distribution. Give one example of each that does not imply the other converse.

Exercise 10.7. Suppose $\sum_n \mathbb{P}(A_n) < \infty$. Prove that with probability one, only finitely many of the events A_n occur.

Exercise 10.8. If $X_n \xrightarrow{\mathbb{P}} c$ where c is a constant, prove that $X_n \xrightarrow{d} c$.

Challenge Exercise 10.9. Let X_n be independent with $\mathbb{P}(X_n = 1) = 1/n$ and $\mathbb{P}(X_n = 0) = 1 - 1/n$. Show that $X_n \xrightarrow{\mathbb{P}} 0$. Use the second Borel–Cantelli lemma to show that X_n does not converge almost surely to 0.

Challenge Exercise 10.10. Suppose $X_n \xrightarrow{a.s.} X$ and $|X_n| \le Y$ for all n, where $\mathbb{E}[Y] < \infty$. Show that $\mathbb{E}[X_n] \to \mathbb{E}[X]$. Explain why this is not a statement about probability convergence alone.

Chapter 11

The Central Limit Theorem and Normal Approximation

The law of large numbers says that averages stabilize. The central limit theorem says something subtler and more informative: after centering by the mean and scaling by the square root of the sample size, the remaining fluctuations are approximately normal. This is one of the main reasons the normal distribution appears everywhere in science.

11.1 From stabilization to fluctuation

The weak and strong laws of large numbers tell us that for i.i.d. random variables $X_1, X_2, \dots$ with mean μ, the sample mean

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

gets close to μ when n is large. That answers one important question: where does the average go? It does *not* answer the next question: how does the average fluctuate before it settles down?

The central limit theorem addresses exactly this issue. It says that when the summands are independent and have finite variance, the random error in the sample mean has a universal shape. After the right centering and scaling, that shape is approximately the standard normal distribution.

This is a remarkable theorem. The original distribution of the X_i can be Bernoulli, exponential, uniform, Poisson, or something much more complicated. Yet the normalized sum tends to the same limiting law.

11.2 Why the square-root scaling appears

Let $S_n = X_1 + \cdots + X_n$, where the X_i are i.i.d. with mean μ and variance $\sigma^2 < \infty$. Then

$$\mathbb{E}[S_n] = n\mu, \qquad \operatorname{Var}(S_n) = n\sigma^2.$$

The mean grows like n, while the standard deviation grows like $\sqrt{n}$. That second fact is the key. It suggests that the natural size of the random fluctuation in S_n is on the order of $\sqrt{n}$, not on the order of n.

Equivalently,

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n},$$

so the sample mean fluctuates on the scale $1/\sqrt{n}$. If we want a nondegenerate limit, we should therefore look at

$$\sqrt{n}(\overline{X}_n - \mu) = \frac{S_n - n\mu}{\sqrt{n}}.$$

Because the natural standard deviation of S_n is $\sigma\sqrt{n}$, it is even more reasonable to standardize further and study

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \qquad \text{or} \qquad \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

This quantity has mean 0 and variance 1 for every n.

11.3 A first example: Bernoulli trials

Suppose $X_i \sim \mathrm{Ber}(p)$ independently, so each X_i takes the value 1 with probability p and 0 with probability $1-p$. Then

$$S_n = X_1 + \cdots + X_n \sim \mathrm{Bin}(n, p),$$

with mean np and variance $np(1-p)$. The quantity

$$\frac{S_n - np}{\sqrt{np(1-p)}}$$

measures how many standard deviations the observed count differs from its mean.

The central limit theorem predicts that for large n, this standardized binomial count should be approximately standard normal. In other words,

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) \approx \Phi(x),$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

is the standard normal cdf.

This approximation explains why even discrete distributions are often treated as approximately normal when the sample size is large.

11.4 The standard normal distribution revisited

The standard normal random variable $Z \sim \mathcal{N}(0,1)$ is characterized by density

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

Its cdf is denoted by Φ. There is no elementary closed form for Φ, but it is well tabulated and built into statistical software.

We record several facts that will matter throughout this chapter.

- The distribution is symmetric about 0, so $\Phi(-x) = 1 - \Phi(x)$.
- The density is bell-shaped, with mean 0 and variance 1.
- If $Z \sim \mathcal{N}(0,1)$, then $a + bZ \sim \mathcal{N}(a, b^2)$ for any real a and any $b > 0$.
- Sums of independent normal random variables are again normal.

The last property is special. Many families of distributions are not closed under convolution. The normal family is, and that algebraic stability is one reason it is so important.

11.5 Statement of the central limit theorem

We now state the classical i.i.d. version.

Theorem 11.1 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2, where $0 < \sigma^2 < \infty$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0,1)$.

This means that for every continuity point x of Φ,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x \right) \to \Phi(x).$$

Since Φ is continuous everywhere, the convergence holds for all real x.

Remark 11.2. The theorem requires finite variance. If the variance is infinite, a different scaling and a different limiting distribution may appear. The normal law is universal, but not completely universal.

11.6 What the theorem does and does not say

The central limit theorem is often paraphrased loosely as "sums become normal." That slogan is helpful, but it can also mislead. Here is the precise picture.

(1) The theorem concerns *centered and scaled* sums, not the raw sums themselves.

(2) The theorem is asymptotic. It says the approximation improves as n grows.

(3) The theorem gives convergence in distribution, not almost sure convergence and not exact equality.

(4) The quality of approximation can vary substantially depending on the skewness, discreteness, or tail behavior of the underlying distribution.

In particular, the theorem is not a license to use normal approximations blindly. One must still ask whether n is large enough relative to the problem at hand.

11.7 A proof under an extra assumption: the mgf method

There are several classical proofs of the central limit theorem. The most general elementary proof uses characteristic functions. An especially transparent proof works under a slightly stronger assumption: that the moment generating function exists in a neighborhood of 0. We present that proof first because it highlights the underlying mechanism cleanly.

Assume for the moment that the i.i.d. variables X_i have mean μ, variance $\sigma^2 > 0$, and moment generating function

$$M_X(t) = \mathbb{E}[e^{tX}]$$

finite for t in some open interval around 0.

Define the standardized variable

$$Y = \frac{X - \mu}{\sigma},$$

so $\mathbb{E}[Y] = 0$ and $\mathrm{Var}(Y) = 1$. Let $Y_1, Y_2, \dots$ be i.i.d. copies of Y, and define

$$T_n = \frac{Y_1 + \cdots + Y_n}{\sqrt{n}}.$$

We want to show $T_n \xrightarrow{d} Z$.

Let $M_Y(t) = \mathbb{E}[e^{tY}]$. Since $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Y^2] = 1$, the Taylor expansion of M_Y around 0 takes the form

$$M_Y(t) = 1 + \frac{t^2}{2} + o(t^2) \qquad \text{as } t \to 0.$$

Now,

$$M_{T_n}(t) = \mathbb{E}\left[e^{t(Y_1+\cdots+Y_n)/\sqrt{n}}\right] = \prod_{j=1}^{n} \mathbb{E}[e^{tY_j/\sqrt{n}}] = \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Using the expansion above,

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right).$$

Hence

$$M_{T_n}(t) = \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{t^2/2},$$

which is the mgf of $\mathcal{N}(0,1)$. By the uniqueness theorem for mgfs,

$$T_n \xrightarrow{d} \mathcal{N}(0,1).$$

Undoing the standardization gives the central limit theorem.

Remark 11.3. This proof is elegant, but the extra mgf assumption excludes some perfectly valid CLT examples, such as certain heavy-tailed distributions with finite variance but no mgf near 0. Characteristic functions solve that problem because they always exist.

11.8 The characteristic-function viewpoint

The most robust proof replaces mgfs by characteristic functions. Let

$$\varphi_Y(t) = \mathbb{E}[e^{itY}].$$

If $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Y^2] = 1$, then one can show

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2) \qquad \text{as } t \to 0.$$

For the normalized sum $T_n = (Y_1 + \cdots + Y_n)/\sqrt{n}$,

$$\varphi_{T_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n \to e^{-t^2/2},$$

which is the characteristic function of the standard normal. Lévy's continuity theorem then implies $T_n \xrightarrow{d} Z$.

This is the standard proof because characteristic functions exist for all distributions, not only those with finite mgfs near 0.

For a first undergraduate course, it is reasonable to present either the mgf proof under a mild extra assumption or the characteristic-function proof in full if characteristic functions have already been developed carefully. What matters most pedagogically is that students see why products of transforms make independent sums tractable.

11.9 The De Moivre–Laplace theorem

The central limit theorem contains as a special case the classical normal approximation to the binomial distribution.

Theorem 11.4 (De Moivre–Laplace)**.** *Let* $S_n \sim \mathrm{Bin}(n,p)$ *with* $0<p<1$*. Then*

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z, \qquad Z \sim \mathcal{N}(0,1).$$

This follows immediately by writing $S_n = X_1 + \cdots + X_n$ with $X_i \sim \mathrm{Ber}(p)$ i.i.d.

The theorem is historically important because the binomial-to-normal approximation was understood before the fully general CLT was proved. Conceptually, it is still important because many practical uses of the CLT begin with sample proportions and count data.

11.9.1 Normal approximation to binomial probabilities

Suppose $S_n \sim \mathrm{Bin}(n,p)$ and we want to approximate

$$\mathbb{P}(a \le S_n \le b).$$

The CLT suggests standardizing:

$$\mathbb{P}(a \le S_n \le b) \approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right).$$

Because the binomial is discrete and the normal is continuous, a correction often improves the approximation.

11.10 Continuity correction

The normal distribution spreads mass continuously over intervals, while the binomial distribution places point mass at integers. To approximate a single value such as $\mathbb{P}(S_n = k)$, it is better to match the integer k with the interval from $k-1/2$ to $k+1/2$.

For example,

$$\mathbb{P}(S_n \le k) = \mathbb{P}(S_n \le k+0.5) \approx \Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right).$$

Similarly,

$$\mathbb{P}(a \le S_n \le b) \approx \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

This is called the *continuity correction*. It is often important when n is moderate rather than huge.

Example 11.5. Suppose $S_{100} \sim \mathrm{Bin}(100, 0.4)$ and we want to approximate $\mathbb{P}(S_{100} \le 35)$. The

mean is 40 and the variance is 24, so the standard deviation is $\sqrt{24} \approx 4.899$.

Without continuity correction,

$$\mathbb{P}(S_{100} \le 35) \approx \Phi\left(\frac{35 - 40}{\sqrt{24}}\right) = \Phi(-1.021) \approx 0.154.$$

With continuity correction,

$$\mathbb{P}(S_{100} \le 35) \approx \Phi\left(\frac{35.5 - 40}{\sqrt{24}}\right) = \Phi(-0.919) \approx 0.179.$$

The second approximation is usually better because it respects the discrete nature of the original problem.

11.11 Approximation for sample means

The CLT is often used through the sample mean rather than the sum. If $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2, then for large n,

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

in the sense of approximate distributional behavior.

Equivalently,

$$\mathbb{P}\left(\overline{X}_n \le x\right) \approx \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right).$$

This is the normal approximation underlying a great deal of classical statistics.

11.11.1 Example: averaging exponentials

Suppose $X_1, \dots, X_n$ are i.i.d. exponential with rate λ. Then $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$. The exact distribution of the sample mean is Gamma-based, but for large n the CLT gives

$$\overline{X}_n \approx \mathcal{N}\left(\frac{1}{\lambda}, \frac{1}{n\lambda^2}\right).$$

Even though the exponential distribution is strongly skewed, the average becomes nearly normal when many terms are combined.

11.12 Confidence-interval intuition

Although this course is about probability rather than inference, it is worth seeing why the CLT is foundational for statistics. If μ and σ are known and $X_1, \dots, X_n$ are i.i.d. with mean μ and

variance σ^2, the CLT suggests that

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx Z.$$

Therefore,

$$\mathbb{P}\left(-1.96 \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95.$$

Rearranging,

$$\mathbb{P}\left(\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95.$$

This is the prototype of a confidence interval. The probability course need not develop the full inference machinery, but students should understand that the CLT is the probability theorem making this kind of interval possible.

11.13 A worked example with sample proportions

Let $X_i \sim \text{Ber}(p)$ be i.i.d. Then

$$\hat{p}_n = \overline{X}_n$$

is the sample proportion of successes. Since $\mu = p$ and $\sigma^2 = p(1-p)$, the CLT gives

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} Z.$$

Thus, for large n,

$$\hat{p}_n \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

This approximation justifies the familiar rule that the standard error of a sample proportion is roughly

$$\sqrt{\frac{p(1-p)}{n}}.$$

When p is unknown, one often plugs in $\hat{p}_n$ for p in applications.

11.14 When the approximation is good

The CLT is asymptotic, so it is natural to ask how large n must be before the normal approximation is usable. There is no universal answer, but several guidelines are helpful.

(1) If the underlying distribution is symmetric and not too heavy-tailed, the approximation is often good surprisingly early.

(2) If the distribution is strongly skewed, more observations may be needed.

(3) For Bernoulli or binomial data, a common rule of thumb is that np and $n(1-p)$ should both be

comfortably larger than 1, and preferably larger than 5 or 10, before a normal approximation is trusted.

(4) If the tails are extremely heavy, the finite-variance assumption may fail, in which case the classical CLT does not apply at all.

These are only heuristics. In serious work, the suitability of the approximation depends on how accurate the answer must be.

11.15 A quantitative bound: the Berry–Esseen theorem

The CLT says the approximation error goes to 0, but not how fast. A famous quantitative refinement is the Berry–Esseen theorem.

Theorem 11.6 (Berry–Esseen, informal statement)**.** *Let $X_1, X_2, \dots$ be i.i.d. with mean μ, variance $\sigma^2 > 0$, and finite third absolute moment $\mathbb{E}|X_1 - \mu|^3 < \infty$. Then there exists an absolute constant C such that for all n,*

$$\sup_x \left| \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x \right) - \Phi(x) \right| \le C \, \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3 \sqrt{n}}.$$

We will not prove this result, but its meaning is important. Under a finite third-moment assumption, the approximation error is typically of order $1/\sqrt{n}$. This helps explain why the CLT becomes useful at moderate sample sizes but is not exact.

11.16 Why the theorem is surprising

It is worth pausing over just how much information the CLT compresses.

Suppose X_i are i.i.d. with mean 10 and variance 4. The theorem says that no matter what the rest of the distribution looks like, provided the variance is finite, the normalized sum behaves approximately normally. The detailed shape of the individual distribution affects the mean and variance, but in the limit it does not affect the normal form of the fluctuations.

This is a universality phenomenon. In mathematics, universality means that large-scale behavior becomes insensitive to many small-scale details. The CLT is one of the earliest and most powerful examples of this idea.

11.17 Lindeberg and beyond

The i.i.d. version is only the beginning. More general central limit theorems allow independent but non-identically distributed summands, triangular arrays, martingale differences, weak dependence, and even some Markovian settings. The key theme remains the same: when many small contributions combine, normalized fluctuations often become approximately Gaussian.

For a first course, it is enough to know that the theorem is robust and extends well beyond the simplest setting. This is one reason Gaussian approximations appear throughout applied probability and statistics.

11.18 Comparison with the law of large numbers

The law of large numbers and the central limit theorem are complementary.

- The law of large numbers says
$$\overline{X}_n \to \mu,$$
so the average becomes predictable.
- The central limit theorem says
$$\sqrt{n}(\overline{X}_n - \mu)$$
has an approximately normal distribution, so the residual randomness after centering is quantified.

The law of large numbers gives the limit. The CLT gives the scale and shape of the error around that limit.

11.19 A useful derivation: tail probabilities for the sample mean

Suppose $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2. For a number $a > 0$,

$$\mathbb{P}(\overline{X}_n - \mu \le a) = \mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le \frac{a\sqrt{n}}{\sigma}\right) \approx \Phi\left(\frac{a\sqrt{n}}{\sigma}\right).$$

Likewise,

$$\mathbb{P}(|\overline{X}_n - \mu| \le a) \approx \Phi\left(\frac{a\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{a\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{a\sqrt{n}}{\sigma}\right) - 1.$$

This shows clearly how larger sample sizes make the sample mean concentrate more tightly around the true mean.

11.20 The standardized sum versus the exact distribution

It is tempting to identify the standardized sum itself with a normal random variable when n is large, but one should remember that the CLT only gives approximate distributional equality.

For example, if $X_i \sim \mathrm{Ber}(1/2)$, then S_n is integer-valued and

$$\frac{S_n - n/2}{\sqrt{n}/2}$$

can only take finitely many values for each fixed n. It is never literally normal. The theorem says only that its cdf becomes close to the normal cdf as n increases.

This distinction may sound pedantic, but it matters. Probability is full of approximations that become powerful precisely because we understand what type of approximation they are.

11.21 A sketch of the normal approximation to the Poisson

If $X \sim \text{Pois}(\lambda)$ and λ is large, then X can also be approximated by a normal random variable with mean λ and variance λ. The heuristic reason is that a Poisson random variable with parameter λ can be represented approximately as a sum of many rare Bernoulli indicators with total mean λ, and the CLT then suggests a Gaussian shape when λ is large.

Thus,

$$\mathbb{P}(X \le k) \approx \Phi\left(\frac{k + 0.5 - \lambda}{\sqrt{\lambda}}\right).$$

As with the binomial case, a continuity correction is appropriate.

11.22 Summary

The central limit theorem is one of the cornerstones of probability.

- It concerns centered and scaled sums of independent random variables.
- The correct fluctuation scale is $\sqrt{n}$.
- Under finite variance,
$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$
- This yields normal approximations for sums, sample means, and sample proportions.
- The continuity correction improves normal approximations to discrete distributions.
- Quantitative refinements such as Berry–Esseen explain how fast the approximation improves.

In the next chapter we study another asymptotic phenomenon with a very different flavor: the emergence of Poisson laws from rare events.

Exercises

Exercise 11.1. Let $X_1, \dots, X_n$ be i.i.d. with mean μ and variance σ^2. Write down the standardized form of the sample mean that appears in the central limit theorem.

Exercise 11.2. Suppose $X_i \sim \text{Ber}(p)$ independently. Show that

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Interpret this result in terms of sample proportions.

Exercise 11.3. Suppose $X_1, \ldots, X_n$ are i.i.d. with mean 5 and variance 9. Use the CLT to approximate

$$\mathbb{P}\left(\overline{X}_n \leq 5.6\right)$$

when $n = 100$.

Exercise 11.4. Let $S_{200} \sim \text{Bin}(200, 0.3)$. Use the normal approximation with continuity correction to approximate $\mathbb{P}(50 \leq S_{200} \leq 70)$.

Exercise 11.5. Explain in words the difference between the law of large numbers and the central limit theorem.

Exercise 11.6. Suppose $X_i \sim \text{Exp}(2)$ independently. What are the mean and variance of $\overline{X}_n$? What normal approximation does the CLT suggest for large n?

Exercise 11.7. Why is the square-root scaling natural in the central limit theorem? Answer using variance.

Exercise 11.8. Let $X \sim \text{Pois}(100)$. Use a normal approximation with continuity correction to estimate $\mathbb{P}(X \leq 110)$.

Exercise 11.9. Assume $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. Show that

$$\text{Var}(\sqrt{n}(\overline{X}_n - \mu)) = \sigma^2.$$

Why is this consistent with the CLT?

Exercise 11.10. State carefully what it means for

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z.$$

Your answer should be written in terms of cdfs.

Challenge Exercise 11.11. Suppose X_i are i.i.d. with mgf finite near 0. Fill in the details of the mgf proof of the CLT by justifying the expansion

$$M_Y(t) = 1 + \frac{t^2}{2} + o(t^2) \quad \text{as } t \to 0$$

for the standardized variable $Y = (X - \mu)/\sigma$.

Challenge Exercise 11.12. Let X_i be i.i.d. with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Use the CLT to approximate

$$\mathbb{P}(S_{400} \geq 20)$$

where $S_n = X_1 + \cdots + X_n$. Compare your approximation to the symmetry of the exact distribution.

Chapter 12

Poisson Approximation and Rare Events

The normal approximation describes the cumulative effect of many moderate fluctuations. Poisson approximation describes a different regime: many opportunities for an event, each one individually unlikely, but with a nontrivial total expected count. In that rare-event world, counts often look Poisson.

12.1 Why a second asymptotic regime is needed

The central limit theorem concerns sums of many independent contributions whose individual effects are not vanishingly small relative to the natural fluctuation scale. There is another basic asymptotic regime that occurs constantly in applications:

- a system has many components, users, sites, or time slots;
- the event of interest is rare for each individual opportunity;
- the total number of opportunities is large enough that a few occurrences are still expected.

Examples are everywhere: the number of defects in a long roll of material, the number of typing errors in a page, the number of calls arriving during a very short interval, the number of network failures in a day, or the number of mutations at a specific site across many replications.

In such settings the relevant limit is often Poisson rather than normal.

12.2 The law of small numbers

The basic result begins with binomial random variables.

Theorem 12.1 (Poisson limit for rare Bernoulli events)**.** *Let* $X_n \sim \mathrm{Bin}(n, p_n)$*, and suppose that*

$$p_n \to 0, \qquad np_n \to \lambda \in (0, \infty).$$

Then for each fixed nonnegative integer k,

$$\mathbb{P}(X_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

Equivalently,

$$X_n \xrightarrow{d} \text{Pois}(\lambda).$$

This is sometimes called the *law of rare events*. It says that a binomial count of many tiny-probability events converges to a Poisson distribution when the expected count remains finite.

Proof. Write

$$\mathbb{P}(X_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k}.$$

For fixed k,

$$\binom{n}{k} p_n^k n = \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k = \frac{(np_n)^k}{k!} \prod_{j=0}^{k-1} \left(1 - \frac{j}{n}\right) \to \frac{\lambda^k}{k!}.$$

Also,

$$(1-p_n)^n \to e^{-\lambda}$$

because $np_n \to \lambda$ and $p_n \to 0$. Finally,

$$(1-p_n)^{-k} \to 1,$$

so

$$(1-p_n)^{n-k} = (1-p_n)^n (1-p_n)^{-k} \to e^{-\lambda}.$$

Combining these limits gives

$$\mathbb{P}(X_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

□

12.3 Heuristic interpretation

The theorem says that when n is large and p is small with $np \approx \lambda$, the exact count of successes in n independent trials behaves roughly like a Poisson random variable of mean λ.

This is reasonable for several intuitive reasons.

(1) The expected number of successes is $np \approx \lambda$.

(2) Because p is very small, it is unlikely that multiple successes are forced to occur together by local crowding effects.

(3) The trials are independent, so rare successes occur in a scattered way.

The Poisson law is therefore a natural limit for sparse random occurrence.

12.4 A transform proof

The probability generating function gives another clean proof. If $X_n \sim \text{Bin}(n, p_n)$, then

$$G_{X_n}(s) = \mathbb{E}[s^{X_n}] = (1 - p_n + p_n s)^n.$$

Rewrite as

$$G_{X_n}(s) = \left(1 + p_n(s-1)\right)^n.$$

Since $p_n \to 0$ and $np_n \to \lambda$,

$$\left(1 + p_n(s-1)\right)^n \to e^{\lambda(s-1)},$$

which is exactly the pgf of $\text{Pois}(\lambda)$. Hence $X_n \xrightarrow{d} \text{Pois}(\lambda)$.

This proof is short, algebraic, and quite memorable. It also extends easily to some more general settings.

12.5 When Poisson approximation is appropriate

Suppose $X \sim \text{Bin}(n, p)$. There are two classical approximations:

- if n is large and p is not too close to 0 or 1, a normal approximation may be appropriate;
- if n is large and p is small while np is moderate, a Poisson approximation may be more natural.

There is no sharp dividing line, but the following informal guide is useful.

- Use Poisson when the event is genuinely rare and the focus is on small counts.
- Use normal when the expected count and the expected failure count are both reasonably large and one is interested in moderate-to-large scale fluctuations around the mean.

For example, if $n = 1000$ and $p = 0.002$, then $np = 2$. A Poisson(2) approximation is very natural. If instead $n = 1000$ and $p = 0.4$, then Poisson is inappropriate while normal is useful.

12.6 Poisson approximation to the binomial

When $X \sim \text{Bin}(n, p)$ with small p, one often uses

$$X \approx \text{Pois}(\lambda), \qquad \lambda = np.$$

This gives immediately

$$\mathbb{P}(X = k) \approx e^{-np} \frac{(np)^k}{k!}$$

for small integers k.

Example 12.2. Suppose a communication system transmits 5000 packets, each of which is corrupted independently with probability 0.0004. Then $X \sim \mathrm{Bin}(5000, 0.0004)$ and $\lambda = np = 2$. The probability of exactly three corrupted packets is approximately

$$\mathbb{P}(X = 3) \approx e^{-2}\frac{2^3}{3!} = \frac{4}{3}e^{-2}.$$

The exact binomial formula is available but more cumbersome, and the Poisson approximation is typically excellent here.

12.6.1 Zero counts

One especially common use is the approximation

$$\mathbb{P}(X = 0) = (1 - p)^n \approx e^{-np}.$$

This is simply the $k = 0$ case of the Poisson approximation, and it appears constantly in reliability, queueing, genetics, and combinatorics.

12.7 Poisson-binomial sums

The approximation extends beyond equal success probabilities. Suppose $I_1, \dots, I_n$ are independent Bernoulli indicators with

$$\mathbb{P}(I_j = 1) = p_j,$$

not necessarily equal, and define

$$W = I_1 + \cdots + I_n.$$

Then

$$\mathbb{E}[W] = \lambda = \sum_{j=1}^{n} p_j.$$

If all the p_j are small, then W is often approximately Poisson(λ).

The pgf gives the heuristic quickly:

$$G_W(s) = \prod_{j=1}^{n} (1 - p_j + p_j s).$$

If each p_j is small, then

$$\log G_W(s) = \sum_{j=1}^{n} \log\left(1 + p_j(s-1)\right) \approx (s-1)\sum_{j=1}^{n} p_j = \lambda(s-1),$$

so

$$G_W(s) \approx e^{\lambda(s-1)}.$$

That is the pgf of a Poisson distribution.

12.7.1 A quantitative bound

There is a famous theorem of Le Cam saying, roughly, that the total variation distance between the law of W and the law of $\mathrm{Pois}(\lambda)$ is at most a constant multiple of $\sum p_j^2$. We record a simplified version.

Theorem 12.3 (Le Cam bound, informal version)**.** *If $W = \sum_{j=1}^{n} I_j$ with independent Bernoulli indicators I_j of means p_j, and if $Z \sim \mathrm{Pois}(\lambda)$ with $\lambda = \sum p_j$, then the approximation error is controlled by $\sum p_j^2$. In particular, if all the p_j are tiny and $\sum p_j^2$ is small, the Poisson approximation is accurate.*

We do not prove this theorem here, but it explains mathematically why "many rare and nearly independent indicators" lead to Poisson counts.

12.8 Occupancy and rare boxes

Poisson approximation is closely connected with occupancy problems. Suppose m balls are thrown independently and uniformly into N boxes, and we ask for the number of boxes receiving some unusual pattern of occupancy.

When the probability that any particular box has the target pattern is small but the number of boxes is large, the total count of such boxes is often approximately Poisson.

Example 12.4 (Empty boxes)**.** Throw m balls independently into N boxes. For box j, let I_j be the indicator that box j remains empty. Then

$$\mathbb{P}(I_j = 1) = \left(1 - \frac{1}{N}\right)^m.$$

Hence the expected number of empty boxes is

$$\lambda = N\left(1 - \frac{1}{N}\right)^m.$$

If the emptiness events were independent, the number of empty boxes would be exactly Poisson-like. They are not independent, but when N is large and the target event is sufficiently sparse, Poisson approximation can still be very good.

The example illustrates an important theme: exact independence is enough for Poisson approximation, but weak dependence can sometimes still permit it.

12.9 Birthday collisions

The birthday problem gives one of the best-known rare-event approximations. Let $X_1, \dots, X_n$ be independent and uniformly distributed on $\{1, \dots, N\}$, representing birthdays (ignoring leap years and nonuniformity). We ask for the probability of at least one collision.

Define indicators

$$I_{ij} = \mathbf{1}_{\{X_i = X_j\}}, \qquad 1 \le i < j \le n.$$

Then the number of matching pairs is

$$W = \sum_{1 \le i < j \le n} I_{ij}.$$

Each indicator has mean $1/N$, so

$$\mathbb{E}[W] = \binom{n}{2} \frac{1}{N}.$$

When n is much smaller than $\sqrt{N}$, collisions are rare and W is approximately Poisson with mean $\binom{n}{2}/N$. Therefore,

$$\mathbb{P}(\text{at least one collision}) = \mathbb{P}(W \ge 1) \approx 1 - e^{-\binom{n}{2}/N}.$$

This Poisson approximation explains the famous heuristic threshold $n \approx \sqrt{N}$ for the onset of likely collisions.

12.10 Poisson approximation for pattern counts

A common combinatorial situation is counting rare local patterns in a long sequence of independent trials.

Example 12.5 (Runs of successes)**.** Toss a coin n times, with success probability p. Let W be the number of positions at which a run of, say, three consecutive heads begins. For each starting point j, define the indicator I_j that tosses $j, j+1, j+2$ are all heads. Then

$$\mathbb{E}[I_j] = p^3.$$

So

$$\mathbb{E}[W] \approx np^3.$$

If p is small and the sequence is long, the events are rare. They are not independent because nearby runs overlap, but in a sparse regime the total count can still be approximately Poisson.

This example shows that Poisson approximation is not limited to exact binomial models. It is really about the sparse occurrence of approximately independent local structures.

12.11 Poisson thinning as a rare-event principle

Suppose events occur according to a large population of opportunities, but each opportunity is retained independently with a small probability. Then the retained count is often Poisson if the total expected number retained is moderate.

At a heuristic level, thinning says this: start with many opportunities, filter them independently, and if the retained events are rare and nearly independent, the resulting count behaves like Poisson. We will see a dynamic version of the same idea in the next chapter when we study thinning of Poisson processes.

12.12 Approximating tail probabilities

If $X \approx \text{Pois}(\lambda)$, then

$$\mathbb{P}(X \geq r) \approx \sum_{k=r}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!},$$

and

$$\mathbb{P}(X \leq r) \approx \sum_{k=0}^{r} e^{-\lambda} \frac{\lambda^k}{k!}.$$

When λ is small or moderate, these finite or rapidly convergent sums are easy to compute.

Example 12.6. A manufacturer produces 10,000 items, each defective independently with probability 0.0002. Then $\lambda = np = 2$, so the probability of at least four defects is approximately

$$1 - \sum_{k=0}^{3} e^{-2} \frac{2^k}{k!}.$$

Because the mean is only 2, the Poisson approximation is much more natural here than a normal approximation.

12.13 Poisson versus normal for Poisson random variables

There is an amusing layering of approximations here.

- Rare Bernoulli counts are approximated by Poisson.
- Large-parameter Poisson random variables are in turn approximated by normal.

Thus a binomial with n large and p small may first be approximated by $\text{Pois}(np)$, and if np itself is large, that Poisson variable may be approximated further by $\mathcal{N}(np, np)$.

This chain of approximations is mathematically natural, but one should not apply it mechanically. If the quantity of interest is a small count, Poisson is usually the more faithful approximation.

12.14 A useful exact identity leading to Poisson limits

One reason Poisson limits arise so often is the approximation

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}.$$

This identity underlies not only the binomial-to-Poisson limit but also many asymptotic estimates in combinatorics and stochastic processes.

For instance, if an event occurs independently in each of n slots with probability λ/n, then the probability of no occurrence is exactly

$$\left(1 - \frac{\lambda}{n}\right)^n,$$

which tends to $e^{-\lambda}$. That is precisely the probability that a Poisson(λ) variable equals 0.

12.15 A multidimensional glimpse

Sometimes one counts several types of rare events at once. If the counts are built from largely disjoint or asymptotically independent structures, the limiting vector may have independent Poisson coordinates. While we will not pursue this systematically, it is worth noting because it anticipates the independent-increment structure of the Poisson process.

12.16 Practical warning: dependence can break the approximation

Poisson approximation depends fundamentally on sparsity and on weak enough dependence. If rare events come in clusters, the approximation can fail badly. For example, if failures propagate from one component to nearby components, the total count of failures may be much more variable than a Poisson model predicts.

Thus the Poisson approximation is not a black box. Before using it, ask:

- Are the individual events rare?
- Are they nearly independent, or at least not strongly clustered?
- Is the expected count moderate rather than huge?

If the answers are yes, Poisson is often an excellent model.

12.17 Summary

This chapter developed the rare-event counterpart to the central limit theorem.

- If $X_n \sim \text{Bin}(n, p_n)$ with $p_n \to 0$ and $np_n \to \lambda$, then $X_n \xrightarrow{d} \text{Pois}(\lambda)$.
- More generally, sums of many independent Bernoulli indicators with small success probabilities are often approximately Poisson.
- Poisson approximation is natural for sparse counts, zero counts, collision counts, and local pattern counts.
- The approximation is especially useful when the expected count is moderate and the events do not cluster too strongly.

The next chapter takes the rare-event idea from static counts to a dynamic counting process in time: the Poisson process.

Exercises

Exercise 12.1. Let $X_n \sim \text{Bin}(n, \lambda/n)$. Show directly from the pmf that for each fixed k,

$$\mathbb{P}(X_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

Exercise 12.2. A system contains 2000 components, each failing independently during a day with probability 0.0015. Use a Poisson approximation to estimate the probability of exactly two failures.

Exercise 12.3. Under the same model as the previous exercise, use a Poisson approximation to estimate the probability of no failures.

Exercise 12.4. Suppose $X \sim \text{Bin}(800, 0.002)$. Which approximation is more natural here, Poisson or normal? Explain briefly and compute the approximating parameter(s).

Exercise 12.5. Let $I_1, \dots, I_n$ be independent Bernoulli indicators with success probabilities $p_1, \dots, p_n$. Show that

$$\mathbb{E}[s^{I_1+\cdots+I_n}] = \prod_{j=1}^{n}(1 - p_j + p_j s).$$

Use this to explain heuristically why a Poisson approximation should hold when all p_j are small.

Exercise 12.6. In the birthday model with n people and N equally likely birthdays, compute the expected number of matching pairs. Use the Poisson heuristic to approximate the probability of at least one matching pair.

Exercise 12.7. If $X \sim \text{Pois}(\lambda)$, show that

$$\mathbb{P}(X = 0) = e^{-\lambda}.$$

Explain why this makes Poisson approximation especially useful for zero-count probabilities.

Exercise 12.8. Suppose that in n independent trials, each trial produces a rare event with probability $2/n$. Use the Poisson approximation to estimate the probability of at least three occurrences when n is large.

Exercise 12.9. Let $X_n \sim \mathrm{Bin}(n, p_n)$ with $np_n \to \lambda$ and $p_n \to 0$. Show that $\mathbb{E}[X_n] \to \lambda$ and $\mathrm{Var}(X_n) \to \lambda$. How is this consistent with the limiting Poisson law?

Challenge Exercise 12.10. Suppose $I_1, \dots, I_n$ are independent with $\mathbb{P}(I_j = 1) = p_j$, and let $W = \sum I_j$. Show that

$$\mathrm{Var}(W) = \sum_{j=1}^{n} p_j(1 - p_j).$$

Compare this with the variance of a Poisson(λ) variable where $\lambda = \sum p_j$. Why does this support the Poisson approximation when the p_j are small?

Challenge Exercise 12.11. Consider n tosses of a fair coin and let W be the number of occurrences of the pattern HHH beginning at some position. Compute $\mathbb{E}[W]$. Why is Poisson approximation plausible when the pattern is longer, say length m, and m is large relative to $\log n$?

Chapter 13

The Poisson Process

A Poisson random variable models the number of rare events in a fixed window. A Poisson process models the evolution of such counts over time. It is the basic stochastic model for randomly scattered arrivals with independent increments and a constant average rate.

13.1 From a single count to a counting process

In the previous chapter we saw that the Poisson distribution arises naturally as a limit law for rare-event counts in a fixed time interval or spatial region. Many applications, however, require more than a single count. We want to know how the count evolves over time.

Examples include:

- phone calls arriving at a switchboard;
- radioactive particles detected by a counter;
- customers arriving at a service desk;
- defects appearing along a moving production line;
- requests reaching a server.

In each case we do not merely ask how many arrivals occur in one hour. We ask how many occur by time t, how many occur between times s and t, how long we wait until the next event, and whether counts over disjoint time intervals behave independently.

These questions lead to the Poisson process.

13.2 Definition by increments

A *counting process* is a family $\{N(t) : t \geq 0\}$ such that $N(t)$ counts how many events have occurred by time t. The canonical Poisson process has four defining properties.

Definition 13.1. A counting process $\{N(t) : t \geq 0\}$ is a *Poisson process with rate* $\lambda > 0$ if:

(i) $N(0) = 0$;

(ii) it has independent increments, meaning that for disjoint intervals the corresponding count increments are independent;

(iii) it has stationary increments, meaning that for $0 \leq s < t$, the distribution of $N(t) - N(s)$ depends only on $t - s$;

(iv) for every $h > 0$,
$$N(t+h) - N(t) \sim \text{Pois}(\lambda h).$$

By the fourth property, for every $t \geq 0$,

$$N(t) \sim \text{Pois}(\lambda t).$$

The parameter λ is the average rate of arrivals per unit time.

Remark 13.2. Sometimes the Poisson process is defined through a small-h description:

$$\mathbb{P}(N(h) = 1) = \lambda h + o(h), \qquad \mathbb{P}(N(h) \geq 2) = o(h),$$

combined with independent and stationary increments. This formulation emphasizes the idea that in a tiny time interval, either nothing happens or one event occurs, with the chance of multiple events being negligible of smaller order.

13.3 Immediate consequences

If $N(t) \sim \text{Pois}(\lambda t)$, then
$$\mathbb{E}[N(t)] = \lambda t, \qquad \text{Var}(N(t)) = \lambda t.$$

Thus both the mean and the variance grow linearly in time.

For $0 \leq s < t$,
$$N(t) - N(s) \sim \text{Pois}(\lambda(t-s)),$$

so
$$\mathbb{E}[N(t) - N(s)] = \lambda(t-s), \qquad \text{Var}(N(t) - N(s)) = \lambda(t-s).$$

Because increments over disjoint intervals are independent, the process has a particularly clean temporal structure.

Example 13.3. If a call center receives calls according to a Poisson process of rate 6 per hour, then the number of calls in two hours is Poisson with mean 12, and the number of calls between 1:15pm and 1:45pm is Poisson with mean 3.

13.4 Small-interval intuition

The rate λ is best understood through very short time intervals. For small h,

$$\mathbb{P}(N(h) = 0) = e^{-\lambda h} = 1 - \lambda h + o(h),$$

$$\mathbb{P}(N(h) = 1) = e^{-\lambda h}(\lambda h) = \lambda h + o(h),$$

and

$$\mathbb{P}(N(h) \geq 2) = 1 - \mathbb{P}(N(h) = 0) - \mathbb{P}(N(h) = 1) = o(h).$$

So, over a tiny interval of length h,

- one arrival occurs with probability about λh;
- no arrival occurs with probability about $1 - \lambda h$;
- two or more arrivals are negligibly unlikely compared with h.

This makes the Poisson process the natural continuous-time analogue of a sequence of independent Bernoulli trials with success probability proportional to the step size.

13.5 Construction from exponential waiting times

There is another, equally important way to define the Poisson process: through interarrival times. This viewpoint is often more intuitive.

Let $T_1, T_2, \ldots$ be i.i.d. exponential random variables with rate λ. Define arrival times

$$S_n = T_1 + \cdots + T_n, \qquad n \geq 1,$$

and set

$$N(t) = \max\{n \geq 0 : S_n \leq t\},$$

with the convention that the maximum of the empty set is 0.

Thus T_1 is the waiting time for the first arrival, T_2 is the waiting time between the first and second arrivals, and so on.

Theorem 13.4. *If $T_1, T_2, \ldots$ are i.i.d. $\text{Exp}(\lambda)$ and $N(t)$ is defined as above, then $\{N(t) : t \geq 0\}$ is a Poisson process with rate λ.*

We will not prove every detail here, but we will verify the most important relationships.

13.5.1 The first arrival time

Let $S_1 = T_1$. Then for $t \geq 0$,

$$\mathbb{P}(S_1 > t) = \mathbb{P}(T_1 > t) = e^{-\lambda t}.$$

Hence the waiting time for the first event is exponential with rate λ.

There is a direct connection with the counting process:

$$\{S_1 > t\} = \{N(t) = 0\}.$$

Therefore,

$$\mathbb{P}(N(t) = 0) = e^{-\lambda t}.$$

That already matches the Poisson(λt) probability of zero arrivals.

13.5.2 The nth arrival time

More generally,

$$\{N(t) \geq n\} = \{S_n \leq t\}.$$

The random variable S_n is a sum of n independent exponential random variables, so it has a Gamma distribution with shape parameter n and rate λ. Its density is

$$f_{S_n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}, \qquad t > 0.$$

The relation between S_n and $N(t)$ is fundamental: arrival times and counting processes are two ways of encoding the same randomness.

13.6 The exponential waiting-time property

One of the signature properties of the Poisson process is the following.

Theorem 13.5. *If $\{N(t)\}$ is a Poisson process of rate λ, then the waiting time until the first arrival is exponential with rate λ.*

Proof. Let W be the waiting time until the first arrival. Then

$$\{W > t\} = \{N(t) = 0\}.$$

Since $N(t) \sim \text{Pois}(\lambda t)$,

$$\mathbb{P}(W > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t}.$$

This is exactly the survival function of an exponential random variable with rate λ. □

More generally, the waiting times between successive arrivals are i.i.d. Exp(λ).

13.7 Memorylessness and the Poisson process

The exponential distribution is the only continuous distribution with the memoryless property:

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t).$$

This property explains why the Poisson process "restarts" after each arrival. Once we have waited s time units without an arrival, the remaining wait has the same distribution as a fresh exponential clock.

That is exactly what one expects from a process with stationary independent increments. The future depends only on the amount of future time, not on how long we have already waited.

13.8 Independent increments from exponential clocks

Why should the interarrival construction produce independent increments? Intuitively, because once one conditions on the current time, the future waiting times are built from fresh independent exponential variables. The memoryless property makes the future of the process probabilistically identical to a new Poisson process started from the current time.

This is one of the earliest examples of a Markovian structure in continuous time.

13.9 Conditional distribution of arrival times given the count

A beautiful property of the Poisson process is that if we condition on the number of arrivals in a fixed interval, then the arrival times are distributed like ordered independent uniforms.

Theorem 13.6. *Condition on the event $\{N(t) = n\}$. Then the unordered arrival times in $(0, t)$ are distributed like n independent $\mathrm{Unif}(0, t)$ random variables, and the ordered arrival times are distributed like the order statistics of those uniforms.*

This theorem is worth remembering because it connects the Poisson process to several earlier topics: conditioning, order statistics, and joint densities.

13.9.1 Why this is plausible

If exactly n arrivals occur in $(0, t)$ and the process has stationary independent increments, then no subinterval should be favored beyond its length. The arrivals should be scattered uniformly over the interval, subject only to the total count being fixed. That is exactly what the uniform-order-statistics statement formalizes.

Example 13.7. Suppose $N(10) = 3$. Then conditional on this event, the three arrival times in $(0, 10)$ have the same joint distribution as the sorted values of three independent uniform random

variables on $(0, 10)$. In particular, the first arrival time is distributed like the minimum of three uniforms.

This result is extremely useful in applied modeling and simulation.

13.10 Merging and splitting Poisson streams

The Poisson process has remarkable closure properties.

13.10.1 Superposition

Suppose $N_1(t)$ and $N_2(t)$ are independent Poisson processes with rates λ_1 and λ_2. Define

$$N(t) = N_1(t) + N_2(t).$$

Then for each fixed t,

$$N(t) \sim \text{Pois}((\lambda_1 + \lambda_2)t)$$

because sums of independent Poisson random variables are Poisson. In fact, the entire process $\{N(t)\}$ is a Poisson process with rate $\lambda_1 + \lambda_2$.

Thus independent Poisson streams add by adding rates.

Example 13.8. Calls arrive from two independent sources at rates 4 and 7 per hour. Then the combined call stream is Poisson with rate 11 per hour.

13.10.2 Thinning

Now suppose $N(t)$ is a Poisson process of rate λ, and each arrival is independently labeled "type A" with probability p and "type B" with probability $1 - p$. Let $N_A(t)$ be the number of type A arrivals by time t and $N_B(t)$ the number of type B arrivals.

Then:

- N_A is a Poisson process with rate λp;
- N_B is a Poisson process with rate $\lambda(1 - p)$;
- N_A and N_B are independent processes.

This is called *thinning*. It is the dynamic analogue of the rare-event filtering idea from the previous chapter.

Example 13.9. Emails arrive to a server according to a Poisson process of rate 30 per hour. Each email is spam independently with probability 0.2. Then spam emails arrive according to a Poisson process of rate 6 per hour.

13.11 The distribution of interarrival times

Let $T_n = S_n - S_{n-1}$ with $S_0 = 0$. Then the T_n are i.i.d. $\text{Exp}(\lambda)$. Therefore,

$$\mathbb{E}[T_n] = \frac{1}{\lambda}, \qquad \text{Var}(T_n) = \frac{1}{\lambda^2}.$$

So the average time between arrivals is $1/\lambda$.

This reciprocal relationship between rate and mean waiting time is basic. If events occur on average λ times per unit time, then the average wait between events is $1/\lambda$ time units.

13.12 The Gamma law for the nth arrival time

Because $S_n = T_1 + \cdots + T_n$ is a sum of n i.i.d. exponentials,

$$S_n \sim \Gamma(n, \lambda)$$

with density

$$f_{S_n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}, \qquad t > 0.$$

Hence

$$\mathbb{E}[S_n] = \frac{n}{\lambda}, \qquad \text{Var}(S_n) = \frac{n}{\lambda^2}.$$

This tells us that the expected time until the nth arrival is n/λ.

Example 13.10. If customers arrive at rate 5 per minute, then the expected time until the tenth arrival is $10/5 = 2$ minutes.

13.13 Covariance structure

For $0 \le s \le t$,

$$N(t) = N(s) + \big(N(t) - N(s)\big),$$

where the increment is independent of $N(s)$. Therefore,

$$\text{Cov}(N(s), N(t)) = \text{Var}(N(s)) = \lambda s.$$

So the covariance is

$$\text{Cov}(N(s), N(t)) = \lambda \min\{s, t\}.$$

This is a simple but important formula. It shows that counts at different times are highly dependent, even though increments over disjoint intervals are independent.

13.14 Conditional expectations in the Poisson process

The independent-increment structure makes conditional expectation very clean. If $0 \le s < t$, then

$$N(t) = N(s) + \big(N(t) - N(s)\big),$$

where the increment is independent of the past and has mean $\lambda(t-s)$. Therefore,

$$\mathbb{E}[N(t) \mid N(s)] = N(s) + \lambda(t-s).$$

In words: given the current count at time s, the best prediction of the future count at time t is the current count plus the expected number of new arrivals in the remaining interval.

This is one of the simplest examples of a conditional-expectation formula in a stochastic process.

13.15 The Poisson process as a continuous-time Markov process

Although we have not yet developed continuous-time Markov chains formally, it is worth noticing that the Poisson process has the Markov property. Given the current count $N(s)$, the future evolution after time s depends on the past only through that present value. Indeed, the future increment process

$$N(s+t) - N(s), \qquad t \ge 0,$$

is independent of the past and has the same law as a fresh Poisson process of rate λ.

This memoryless restart property is one reason the Poisson process is so central in stochastic modeling.

13.16 Simulation

There are two natural ways to simulate a Poisson process on a finite interval $[0, t]$.

13.16.1 Method 1: simulate interarrival times

Generate i.i.d. $\text{Exp}(\lambda)$ waiting times $T_1, T_2, \ldots$, form partial sums S_n, and keep all arrivals with $S_n \le t$.

13.16.2 Method 2: condition on the total count

First generate $N(t) \sim \text{Pois}(\lambda t)$. Then, conditional on $N(t) = n$, generate n i.i.d. $\text{Unif}(0, t)$ points and sort them.

The second method is a direct application of the conditional order-statistics property.

13.17 A small derivation using partitioning

One can motivate the Poisson process by dividing time into many tiny subintervals of equal length $\Delta t = t/m$. In each small subinterval let an arrival occur with probability approximately $\lambda \Delta t$, independently across intervals, and neglect the chance of multiple arrivals in a single tiny interval.

Then the number of arrivals by time t is approximately binomial:

$$N(t) \approx \text{Bin}\left(m, \lambda \frac{t}{m}\right).$$

As $m \to \infty$, the binomial distribution tends to $\text{Pois}(\lambda t)$. This connects the dynamic process directly to the rare-event limits of the previous chapter.

13.18 Applications

The Poisson process is used because it is mathematically tractable and often empirically reasonable over moderate time scales. Here are a few standard interpretations.

- **Queueing.** Customers arrive randomly to a service station.
- **Reliability.** Failures occur over time in a large engineered system.
- **Telecommunications.** Requests or packets arrive at a network node.
- **Physics.** Particle detections or decay events occur randomly in time.
- **Biology.** Mutations or cell arrivals occur at roughly constant average rates.

Of course, real systems need not satisfy the independence or constant-rate assumptions exactly. The Poisson process is a model, not a law of nature. Still, it is often the right starting point.

13.19 When the Poisson process is not appropriate

It is just as important to know when the model fails.

- If arrivals come in bursts or clusters, independent increments may fail.
- If the rate changes systematically over time, stationarity fails; one may need a nonhomogeneous Poisson process.
- If events inhibit one another, the process may be more regular than Poisson.

These failures do not diminish the importance of the Poisson process. They simply show that one must match model assumptions to data and mechanism.

13.20 A brief look at nonhomogeneous Poisson processes

If the rate varies with time, say according to a deterministic intensity function $\lambda(t)$, then one can define a *nonhomogeneous Poisson process* by requiring that increments over disjoint intervals remain independent and that

$$N(t) - N(s) \sim \text{Pois}\left(\int_s^t \lambda(u)\,du\right).$$

The homogeneous Poisson process studied in this chapter is the special case in which $\lambda(t) \equiv \lambda$ is constant.

We will not develop this generalization further, but it is useful to know it exists.

13.21 Summary

The Poisson process is the basic continuous-time counting model.

- It has independent, stationary increments.
- Counts over intervals of length h are Poisson with mean λh.
- Interarrival times are i.i.d. exponential with rate λ.
- The nth arrival time has a Gamma distribution.
- Conditional on $N(t) = n$, the arrival times in $(0, t)$ are uniform order statistics.
- Superposition adds rates; thinning multiplies rates.

The next chapter turns from continuous-time counting to another foundational stochastic model: discrete-time Markov chains.

Exercises

Exercise 13.1. If $N(t)$ is a Poisson process with rate λ, what is the distribution of $N(5) - N(2)$?

Exercise 13.2. Calls arrive according to a Poisson process of rate 8 per hour. Compute the probability of exactly three calls in a half-hour interval.

Exercise 13.3. For a Poisson process of rate λ, show that

$$\mathbb{P}(N(h) = 0) = 1 - \lambda h + o(h)$$

and

$$\mathbb{P}(N(h) = 1) = \lambda h + o(h)$$

as $h \downarrow 0$.

Exercise 13.4. Let W be the waiting time until the first arrival in a rate-λ Poisson process. Compute $\mathbb{P}(W > t)$ and identify the distribution of W.

Exercise 13.5. If arrivals occur at rate 3 per minute, what is the expected time until the fourth arrival?

Exercise 13.6. Suppose N_1 and N_2 are independent Poisson processes with rates 2 and 5. What is the distribution of $N_1(t) + N_2(t)$?

Exercise 13.7. A rate-10 Poisson process is thinned by independently keeping each arrival with probability 0.3. What is the distribution of the retained process?

Exercise 13.8. If $N(t)$ is a Poisson process, compute $\mathrm{Cov}(N(2), N(5))$.

Exercise 13.9. Condition on the event $N(1) = 2$. What is the conditional distribution of the two arrival times in $(0, 1)$?

Exercise 13.10. Show that for $0 \le s < t$,

$$\mathbb{E}[N(t) \mid N(s)] = N(s) + \lambda(t - s).$$

Challenge Exercise 13.11. Let S_n be the time of the nth arrival in a rate-λ Poisson process. Prove that

$$\mathbb{P}(S_n \le t) = \mathbb{P}(N(t) \ge n).$$

Explain this identity in words.

Challenge Exercise 13.12. Show that conditional on $N(t) = n$, the waiting time until the first arrival has the same distribution as the minimum of n i.i.d. $\mathrm{Unif}(0, t)$ random variables. Compute its density.

Chapter 14

Markov Chains

A Markov chain is a stochastic process whose future depends on the present state but, given the present, not on the more distant past. This "memoryless given the present" structure is simple enough to analyze and rich enough to model a huge range of random systems.

14.1 Why Markov chains matter

Many random systems evolve step by step. A customer's location in a queueing network changes from one time period to the next. A board-game token moves from square to square. A website user's status changes from inactive to active and back again. A population can move between genetic states. A random walk jumps from vertex to vertex on a graph.

What makes these examples analytically tractable is often the same structural simplification: once we know the current state, the future evolution no longer depends on the full previous history. That is the idea behind a Markov chain.

The topic is important for at least three reasons.

(1) It gives one of the cleanest introductions to stochastic processes.

(2) It unifies probability, linear algebra, and recursion.

(3) It underlies many modern applications, from search algorithms and Monte Carlo methods to genetics, epidemiology, and queueing.

14.2 Definition and transition probabilities

Let S be a finite or countable set, called the *state space*. A sequence of random variables

$$X_0, X_1, X_2, \dots$$

taking values in S is a *Markov chain* if for every $n \geq 0$ and all states $i_0, \dots, i_n, j \in S$,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i_n)$$

whenever the conditioning event has positive probability.

The common value is denoted

$$p_{i_n j} = \mathbb{P}(X_{n+1} = j \mid X_n = i_n).$$

If these one-step transition probabilities do not depend on n, the chain is called *time-homogeneous*. In this chapter, unless stated otherwise, "Markov chain" means time-homogeneous Markov chain.

Definition 14.1. For a time-homogeneous Markov chain, the matrix

$$P = (p_{ij})_{i,j \in S}$$

is called the *transition matrix*. It satisfies

$$p_{ij} \geq 0, \qquad \sum_{j \in S} p_{ij} = 1$$

for every state i.

A matrix with nonnegative entries whose row sums equal 1 is called *stochastic* or *row-stochastic*. Each row describes a probability distribution for the next state given the current state.

14.3 Examples

14.3.1 A two-state weather model

Let $S = \{R, S\}$ for rainy and sunny. Suppose

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

Then if today is rainy, tomorrow is rainy with probability 0.7 and sunny with probability 0.3. If today is sunny, tomorrow is sunny with probability 0.6.

This is one of the simplest nontrivial Markov chains, but it already illustrates many basic ideas: long-run behavior, stationary distributions, and matrix powers.

14.3.2 Simple random walk on the integers

Let $S = \mathbb{Z}$, and suppose

$$\mathbb{P}(X_{n+1} = i + 1 \mid X_n = i) = p, \qquad \mathbb{P}(X_{n+1} = i - 1 \mid X_n = i) = 1 - p.$$

This is the simple random walk. The chain is homogeneous and spatially symmetric when $p = 1/2$.

14.3.3 Inventory model

Suppose a store's inventory level is checked daily. The state is the inventory level at the end of each day, truncated at some maximum capacity. If the replenishment rule depends only on today's stock level and demand is independent from day to day, the resulting sequence is often Markovian.

14.3.4 Board games

A token moving around a finite board according to dice rolls produces a Markov chain. The state is the current square. Special squares that send the token elsewhere simply change the transition probabilities.

14.4 The Markov property in words

The Markov property is often summarized by saying that the future is independent of the past given the present. That phrase is accurate but should be interpreted carefully.

- The future is *not* independent of the past unconditionally.
- The present state may itself encode information from the past.
- What the property says is that once the present state is known, the more distant history provides no further help in predicting the next step.

This balance between memorylessness and structure is exactly what makes Markov chains powerful.

14.5 Initial distribution and evolution

To specify a Markov chain completely, we need both the transition matrix P and the distribution of the starting state X_0.

If the initial distribution is

$$\mu_i = \mathbb{P}(X_0 = i),$$

then the distribution of X_1 is

$$\mathbb{P}(X_1 = j) = \sum_i \mu_i p_{ij}.$$

In vector notation, if we treat μ as a row vector, then

$$\mu^{(1)} = \mu P.$$

Likewise, the distribution of X_n is

$$\mu^{(n)} = \mu P^n.$$

Thus the evolution of state distributions is governed by powers of the transition matrix.

14.6 n-step transition probabilities

Define

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i).$$

These are the *n-step transition probabilities*. They form the entries of the matrix P^n.

Theorem 14.2. *For every $n \geq 1$,*

$$P^n = (p_{ij}^{(n)})_{i,j \in S}.$$

Proof. The case $n = 1$ is the definition. Assume the result for n. Then

$$p_{ij}^{(n+1)} = \mathbb{P}(X_{n+1} = j \mid X_0 = i).$$

Condition on the intermediate state $X_n = k$:

$$p_{ij}^{(n+1)} = \sum_k \mathbb{P}(X_{n+1} = j \mid X_n = k, X_0 = i)\mathbb{P}(X_n = k \mid X_0 = i).$$

By the Markov property,

$$\mathbb{P}(X_{n+1} = j \mid X_n = k, X_0 = i) = p_{kj},$$

so

$$p_{ij}^{(n+1)} = \sum_k p_{ik}^{(n)} p_{kj}.$$

This is exactly the (i, j) entry of $P^n P = P^{n+1}$. □

14.7 Chapman–Kolmogorov equations

The identity used in the proof above is fundamental.

Theorem 14.3 (Chapman–Kolmogorov equations)**.** *For all $m, n \geq 0$ and all states i, j,*

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)}.$$

This formula simply says that to go from i to j in $m + n$ steps, the chain must pass through some intermediate state k after m steps and then go from k to j in the next n steps.

14.8 Communication and accessibility

The long-run structure of a chain depends strongly on which states can lead to which other states.

Definition 14.4. We say that state j is *accessible* from state i, written $i \to j$, if there exists $n \geq 0$ such that

$$p_{ij}^{(n)} > 0.$$

States i and j *communicate*, written $i \leftrightarrow j$, if $i \to j$ and $j \to i$.

Communication is an equivalence relation, so it partitions the state space into *communicating classes*.

Definition 14.5. A Markov chain is *irreducible* if every pair of states communicates.

In an irreducible chain, the entire state space is a single communicating class.

Example 14.6. A simple random walk on a finite cycle graph is irreducible because from any state one can reach any other state by taking enough clockwise or counterclockwise steps.

14.9 Closed classes and absorbing states

Some classes, once entered, cannot be left.

Definition 14.7. A subset $C \subseteq S$ is *closed* if whenever $i \in C$ and $p_{ij} > 0$, we also have $j \in C$.

Equivalently, from states in C the chain cannot jump outside C in one step, and hence cannot leave C at all.

Definition 14.8. A state i is *absorbing* if

$$p_{ii} = 1.$$

An absorbing state forms a closed class by itself.

Example 14.9. In gambler's ruin with fortunes $0, 1, \ldots, N$, the states 0 and N are absorbing because once the gambler is broke or has reached the target fortune, the process stops.

14.10 Hitting times and return times

A central probabilistic question is whether the chain ever hits a given state or set.

Definition 14.10. For a subset $A \subseteq S$, the *hitting time* of A is

$$\tau_A = \inf\{n \geq 0 : X_n \in A\}.$$

For a state i, the first return time to i is

$$T_i^+ = \inf\{n \geq 1 : X_n = i\}.$$

These random times may be finite or infinite. Their behavior separates recurrent from transient states.

14.11 Recurrence and transience

Definition 14.11. A state i is *recurrent* if

$$\mathbb{P}_i(T_i^+ < \infty) = 1,$$

meaning that starting from i, the chain returns to i with probability 1. Otherwise i is *transient*.

Here $\mathbb{P}_i$ means probability for the chain started from state i.

Recurrence means that the state is revisited eventually with certainty. Transience means there is a positive probability that once the chain leaves the state, it never comes back.

Theorem 14.12. *In an irreducible Markov chain, either every state is recurrent or every state is transient.*

We will not prove this theorem fully here, but the intuition is straightforward: in an irreducible chain, all states communicate, so the chain's long-run tendency to return or drift away cannot differ from one state to another.

14.11.1 Finite irreducible chains are recurrent

A particularly important special case is finite state space.

Theorem 14.13. *Every finite irreducible Markov chain has all states recurrent.*

Idea of proof. Because the state space is finite, the chain cannot keep visiting new states forever. In an irreducible chain, every state remains reachable from every other state. One can show that some state must be visited infinitely often with probability 1, and irreducibility then forces recurrence of all states. □

This theorem has an important conceptual message: on a finite connected state space, a chain cannot permanently wander away.

14.12 Periodicity

A recurrent state may still be visited only at multiples of some fixed period.

Definition 14.14. The *period* of a state i is

$$d(i) = \gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}.$$

If $d(i) = 1$, the state is *aperiodic*. Otherwise it is periodic.

Example 14.15. Consider the simple random walk on the even-odd line $\mathbb{Z}$ with nearest-neighbor jumps. Starting from 0, the chain can return to 0 only after an even number of steps. Hence the period is 2.

Theorem 14.16. *If states i and j communicate, then they have the same period.*

So in an irreducible chain the period is a class property, and we speak of an irreducible chain as periodic or aperiodic.

14.13 Why periodicity matters

If a chain has period $d > 1$, the distribution at time n oscillates among residue classes modulo d. As a result, the distribution of X_n need not converge as $n \to \infty$, even if the chain is finite and irreducible.

Aperiodicity removes this oscillatory obstruction and is the key condition for convergence to equilibrium in the finite irreducible case.

14.14 Stationary distributions

A probability distribution π on S is *stationary* if

$$\pi P = \pi.$$

In coordinates, this means

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

Definition 14.17. A stationary distribution is a probability vector that is unchanged by one step of the chain.

If $X_0 \sim \pi$ and π is stationary, then for every n,

$$X_n \sim \pi.$$

So stationarity means the chain is already in equilibrium from the start.

14.14.1 Example: two-state chain

For

$$P = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}, \qquad 0 < a, b < 1,$$

a stationary distribution $\pi = (\pi_1, \pi_2)$ solves

$$\pi_1 = \pi_1 a + \pi_2 (1-b), \qquad \pi_1 + \pi_2 = 1.$$

Solving gives

$$\pi_1 = \frac{1-b}{2-a-b}, \qquad \pi_2 = \frac{1-a}{2-a-b}.$$

This is the long-run fraction of time spent in each state when the chain is irreducible and aperiodic.

14.15 Existence of stationary distributions

For finite state spaces, stationary distributions always exist. This can be shown using linear algebra or a compactness argument. In infinite state spaces, stationary distributions may fail to exist.

Theorem 14.18. *Every finite Markov chain has at least one stationary distribution.*

In a finite irreducible chain, the stationary distribution is unique and assigns positive probability to every state.

Theorem 14.19. *A finite irreducible Markov chain has a unique stationary distribution π, and $\pi_i > 0$ for every state i.*

14.16 Detailed balance and reversible chains

Sometimes stationarity can be verified through a stronger condition.

Definition 14.20. A distribution π satisfies the *detailed balance equations* if for all states i, j,

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

If detailed balance holds, then π is stationary, because summing over i gives

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji},$$

and after relabeling and using the balance equation pairwise one gets $\pi P = \pi$.

More directly,

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji},$$

which in finite settings is easily checked to equal π_j when the balance equations are imposed in the usual pairwise way. The real content is that the net probability flow from i to j equals the net flow from j to i.

Chains satisfying detailed balance are called *reversible*. Reversibility is extremely important in random walks on graphs and in Markov chain Monte Carlo, though we will use it here mainly as a computational tool.

14.16.1 Random walk on an undirected graph

Let $G = (V, E)$ be a finite connected undirected graph. A simple random walk chooses uniformly among neighboring vertices at each step. Then

$$p_{ij} = \begin{cases} 1/\deg(i), & \text{if } i \sim j, \\ 0, & \text{otherwise.} \end{cases}$$

A stationary distribution is given by

$$\pi_i = \frac{\deg(i)}{\sum_{v \in V} \deg(v)}.$$

Indeed,

$$\pi_i p_{ij} = \frac{\deg(i)}{\sum_v \deg(v)} \cdot \frac{1}{\deg(i)} = \frac{1}{\sum_v \deg(v)}$$

whenever $i \sim j$, and the same expression holds for $\pi_j p_{ji}$. Thus detailed balance holds.

14.17 First-step analysis

Many questions about Markov chains lead to recursive equations obtained by conditioning on the first step. This method is called *first-step analysis*. It is one of the most useful computational techniques in the subject.

14.17.1 Hitting probabilities

Suppose we want the probability of hitting a target set A before another set B. Define

$$h(i) = \mathbb{P}_i(\tau_A < \tau_B).$$

Then for states in A, $h(i) = 1$, and for states in B, $h(i) = 0$. For states outside $A \cup B$, conditioning on the first step gives

$$h(i) = \sum_j p_{ij} h(j).$$

So the hitting probabilities satisfy a system of linear equations.

14.17.2 Expected hitting times

If we want the expected time to hit a set A, define

$$m(i) = \mathbb{E}_i[\tau_A].$$

Then $m(i) = 0$ for $i \in A$, and for $i \notin A$,

$$m(i) = 1 + \sum_j p_{ij} m(j).$$

The extra 1 accounts for the first step. Again we obtain a linear system.

This is the basic pattern: define the desired quantity, condition on the first step, and solve the resulting equations.

14.18 Example: gambler's ruin

A gambler starts with fortune $i \in \{0, 1, \dots, N\}$ and on each play either gains 1 with probability p or loses 1 with probability $q = 1 - p$. The process stops at 0 or N. This is a Markov chain on $\{0, 1, \dots, N\}$ with absorbing states 0 and N.

14.18.1 Probability of reaching N before 0

Let

$$h(i) = \mathbb{P}_i(\tau_N < \tau_0).$$

Then

$$h(0) = 0, \qquad h(N) = 1,$$

and for $1 \le i \le N - 1$,

$$h(i) = p\,h(i+1) + q\,h(i-1).$$

This second-order difference equation can be solved explicitly.

Theorem 14.21 (Gambler's ruin probabilities)**.** *If $p \neq q$, then*

$$h(i) = \frac{1 - (q/p)^i}{1 - (q/p)^N}.$$

If $p = q = 1/2$, then

$$h(i) = \frac{i}{N}.$$

Sketch of proof. For $p \neq q$, try a solution of the form r^i, leading to the quadratic equation

$$pr^2 - r + q = 0,$$

whose roots are 1 and q/p. Thus the general solution is

$$h(i) = A + B(q/p)^i.$$

Use $h(0) = 0$ and $h(N) = 1$ to determine A and B. For $p = q$, the difference equation becomes

$$h(i) = \frac{1}{2}h(i+1) + \frac{1}{2}h(i-1),$$

whose solutions are linear, so $h(i) = A + Bi$ and the boundary conditions give $h(i) = i/N$. □

14.18.2 Expected duration

Let

$$m(i) = \mathbb{E}_i[\tau_{\{0,N\}}]$$

be the expected number of plays until absorption. Then

$$m(0) = m(N) = 0,$$

and for $1 \le i \le N-1$,

$$m(i) = 1 + pm(i+1) + qm(i-1).$$

In the fair case $p = q = 1/2$, the solution is especially simple.

Theorem 14.22 (Expected duration in the fair case)**.** *If $p = q = 1/2$, then*

$$m(i) = i(N-i).$$

Proof. We need to solve

$$m(i) = 1 + \frac{1}{2}m(i+1) + \frac{1}{2}m(i-1)$$

with $m(0) = m(N) = 0$. Rearranging gives

$$m(i+1) - 2m(i) + m(i-1) = -2.$$

A quadratic polynomial is natural; try $m(i) = ai^2 + bi + c$. Substituting shows $a = -1$, and the boundary conditions then give $c = 0$ and $b = N$. Hence

$$m(i) = i(N-i).$$

□

This is an excellent illustration of first-step analysis in action.

14.19 Absorbing chains and fundamental matrices

A finite Markov chain with at least one absorbing state is called an *absorbing chain* if from every state there is a positive probability of eventually reaching an absorbing state.

If the states are ordered so that transient states come first and absorbing states last, the transition matrix can be written in block form as

$$P = \begin{pmatrix} Q & R \\ 0 & I \end{pmatrix},$$

where Q describes transitions among transient states.

The matrix

$$(I - Q)^{-1} = I + Q + Q^2 + \cdots$$

is called the *fundamental matrix*. Its (i, j) entry gives the expected number of visits to transient state j starting from transient state i.

We will not develop the full absorbing-chain theory in detail, but it is useful to know that matrix methods can encode hitting probabilities and expected absorption times compactly.

14.20 Long-run behavior in the finite irreducible aperiodic case

The central convergence theorem for finite-state Markov chains is the following.

Theorem 14.23 (Convergence to stationarity)**.** *Let P be the transition matrix of a finite irreducible aperiodic Markov chain, and let π be its unique stationary distribution. Then for all states i, j,*

$$p_{ij}^{(n)} \to \pi_j \qquad \text{as } n \to \infty.$$

Equivalently, regardless of the initial distribution,

$$\mu P^n \to \pi.$$

This theorem says that the chain forgets its starting point in distribution and approaches equilibrium.

Remark 14.24. If the chain is irreducible but periodic, the stationary distribution still exists and is unique, but P^n need not converge. The chain may continue to oscillate among phases. Time averages often still converge, but pointwise distribution convergence can fail.

14.21 Interpretation of the stationary distribution

For a finite irreducible chain, the stationary distribution has two related interpretations.

(1) It is the limiting distribution of X_n when the chain is aperiodic.

(2) It describes the long-run fraction of time spent in each state.

Thus if $\pi_i = 0.2$, then over a long run the chain spends about 20% of the time in state i.

This second interpretation remains valid more broadly than the first and connects Markov chains to ergodic ideas.

14.22 Expected return times and stationary probabilities

In a finite irreducible chain there is a deep and useful relation between stationary probability and expected return time.

Theorem 14.25. *Let π be the stationary distribution of a finite irreducible chain. Then the expected return time to state i satisfies*

$$\mathbb{E}_i[T_i^+] = \frac{1}{\pi_i}.$$

This formula is intuitive. If the chain spends a fraction π_i of its time in state i, then the average time between successive visits to i should be about $1/\pi_i$.

We will not prove the theorem here, but it is worth knowing. It often turns a stationary computation into a probabilistic interpretation.

14.23 Aperiodicity by self-loops

A very common way to guarantee aperiodicity is to have a positive chance of staying in place.

Proposition 14.26. *If an irreducible chain has some state i with $p_{ii} > 0$, then the chain is aperiodic.*

Proof. If $p_{ii} > 0$, then $p_{ii}^{(1)} > 0$, so the period of i divides 1 and must therefore be 1. In an irreducible chain all states have the same period, so the whole chain is aperiodic. □

This simple observation is frequently useful in modeling and in numerical algorithms.

14.24 Random walks on finite graphs

Random walk on a graph gives a rich family of examples with clean stationary distributions. Let $G = (V, E)$ be a finite connected undirected graph, and let the chain move at each step to a uniformly chosen neighbor.

We already showed that the stationary distribution is proportional to degree:

$$\pi(v) = \frac{\deg(v)}{\sum_{u \in V} \deg(u)}.$$

If the graph is not bipartite, the walk is aperiodic and hence converges to this distribution. If the graph is bipartite, the walk has period 2 and oscillates between the two parts.

This example is conceptually valuable because it ties graph structure to Markov behavior.

14.25 Conditional expectation and martingale ideas

Markov chains also provide a clean setting for conditional expectation. If f is a real-valued function on the state space, then

$$\mathbb{E}[f(X_{n+1}) \mid X_n = i] = \sum_j p_{ij} f(j).$$

More compactly, defining

$$(Pf)(i) = \sum_j p_{ij} f(j),$$

we have

$$\mathbb{E}[f(X_{n+1}) \mid X_n] = (Pf)(X_n).$$

This operator viewpoint is extremely useful. For example, if $Pf = f$, then $f(X_n)$ is a martingale. If $Pf \le f$, then $f(X_n)$ is a supermartingale. We will return briefly to this connection in the next chapter.

14.26 A finite-state example in full

Consider the chain on $S = \{0, 1, 2\}$ with transition matrix

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

This can be interpreted as a random walk on the line segment 0–1–2 with holding probability.

14.26.1 Irreducibility and aperiodicity

The chain is irreducible because each state can reach every other state through the middle state. It is aperiodic because every state has positive self-loop probability.

14.26.2 Stationary distribution

By symmetry, one expects $\pi_0 = \pi_2$. Solving $\pi P = \pi$ and $\pi_0 + \pi_1 + \pi_2 = 1$ gives

$$\pi = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \right).$$

Thus in the long run the chain spends half its time in the middle state.

14.26.3 Long-run behavior

Because the chain is finite, irreducible, and aperiodic,

$$P^n \to \begin{pmatrix} \pi \\ \pi \\ \pi \end{pmatrix}$$

as $n \to \infty$, meaning that every row of P^n converges to the stationary distribution.

Working through one concrete example like this is good practice because it combines graph structure, matrix algebra, and probabilistic interpretation.

14.27 What to compute in practice

When faced with a new Markov chain, the following checklist is useful.

(1) Identify the state space and the transition matrix.

(2) Determine communicating classes and whether the chain is irreducible.

(3) Check for absorbing states or closed classes.

(4) Determine periodicity.

(5) If the state space is finite and irreducible, solve for the stationary distribution.

(6) Use first-step analysis for hitting probabilities and expected hitting times.

(7) If the chain is finite, irreducible, and aperiodic, conclude convergence to stationarity.

This is the basic workflow for nearly every elementary Markov-chain problem.

14.28 Summary

This chapter introduced the main ideas of discrete-time Markov chains.

- A Markov chain has the property that the future depends on the present state but not, given the present, on the entire past.
- The transition matrix P determines the dynamics, and P^n gives n-step transitions.
- Communication classes describe reachability structure.
- States may be recurrent or transient; finite irreducible chains are recurrent.
- Periodicity affects whether distributions converge in time.

- Stationary distributions satisfy $\pi P = \pi$ and describe equilibrium behavior.
- First-step analysis turns hitting questions into linear equations.
- Finite irreducible aperiodic chains converge to their unique stationary distribution.

The next chapter gives a brief look at selected further topics that extend the main course narrative.

Exercises

Exercise 14.1. Verify that the rows of any transition matrix sum to 1. Why is this necessary probabilistically?

Exercise 14.2. Consider the two-state chain with transition matrix

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}.$$

Find the stationary distribution.

Exercise 14.3. Let X_n be a Markov chain with transition matrix P. Show that the distribution of X_n is μP^n if the initial distribution is μ.

Exercise 14.4. Compute P^2 for the two-state chain in the previous exercise and interpret its entries probabilistically.

Exercise 14.5. For gambler's ruin with $p = q = 1/2$, verify directly that the hitting probability of N before 0 is i/N.

Exercise 14.6. In the fair gambler's ruin chain on $\{0, 1, 2, 3\}$, compute the expected time to absorption starting from state 1.

Exercise 14.7. Show that if a finite irreducible chain has a self-loop at some state, then it is aperiodic.

Exercise 14.8. Consider a simple random walk on the triangle graph with vertices $1, 2, 3$. Find the stationary distribution.

Exercise 14.9. Let $h(i) = \mathbb{P}_i(\tau_A < \tau_B)$. Derive the first-step equation

$$h(i) = \sum_j p_{ij} h(j)$$

for states $i \notin A \cup B$.

Exercise 14.10. Let $m(i) = \mathbb{E}_i[\tau_A]$. Derive the first-step equation

$$m(i) = 1 + \sum_j p_{ij} m(j)$$

for states $i \notin A$.

Exercise 14.11. Show that if π satisfies the detailed balance equations, then π is stationary.

Exercise 14.12. For simple random walk on a finite connected undirected graph, verify the stationary distribution proportional to degree.

Challenge Exercise 14.13. Let P be the transition matrix of a finite irreducible chain and let f be a function on the state space. Define

$$(Pf)(i) = \sum_j p_{ij} f(j).$$

Show that

$$\mathbb{E}[f(X_{n+1}) \mid X_n] = (Pf)(X_n).$$

Then interpret the condition $Pf = f$ probabilistically.

Challenge Exercise 14.14. Consider simple random walk on the four-cycle $\{0, 1, 2, 3\}$ moving to a neighboring vertex with probability $1/2$. Determine whether the chain is irreducible, whether it is periodic, and what the stationary distribution is.

Challenge Exercise 14.15. For gambler's ruin with $p \neq q$, derive the formula

$$h(i) = \frac{1 - (q/p)^i}{1 - (q/p)^N}$$

carefully from the recursion.

Chapter 15

Selected Further Topics: Branching Processes and Martingales

A one-semester probability course cannot cover all of modern probability, but it should leave students with a sense of where the subject goes next. Two natural bridges are branching processes, which model reproduction and growth, and martingales, which formalize the idea of a fair game and organize many powerful arguments in probability.

15.1 Why include a final topics chapter?

The core narrative of this course has already introduced the central structures of undergraduate probability: random variables, conditioning, expectation, laws of large numbers, central limit theorems, the Poisson process, and Markov chains. Yet several ideas have appeared repeatedly in the background without being fully named.

- Generating functions naturally encode the evolution of offspring counts and total population sizes.
- Conditional expectation provides the language for sequential prediction.
- Markov chains and stochastic processes suggest that some functions of the state might have especially simple dynamics.

This chapter develops those themes through two linked topics.

(1) **Branching processes**, which are simple stochastic models of reproduction and extinction.

(2) **Martingales**, which are processes whose conditional expected future value equals the present.

Both topics lie just beyond the standard first-course boundary, but both are accessible with the tools we already have.

15.2 Galton–Watson branching processes

A *Galton–Watson branching process* models a population that evolves generation by generation. Each individual produces a random number of children, independently of other individuals and according to a common offspring distribution.

Definition 15.1. Let ξ be a nonnegative integer-valued random variable, called the *offspring distribution*. A branching process $\{Z_n : n \geq 0\}$ with offspring law ξ is defined by

$$Z_0 \in \{0, 1, 2, \dots\},$$

and, for $n \geq 0$,

$$Z_{n+1} = \sum_{k=1}^{Z_n} \xi_{n,k},$$

where the $\xi_{n,k}$ are i.i.d. copies of ξ, independent across all n and k.

Thus Z_n is the population size in generation n. If $Z_n = 0$, then automatically $Z_{n+1} = 0$, so extinction is absorbing.

15.2.1 Interpretation

The original historical model was intended to study survival of surnames, but the same mathematics appears in epidemiology, cell division, network cascades, nuclear chain reactions, and random tree growth.

15.3 Offspring generating functions

Let

$$f(s) = \mathbb{E}[s^{\xi}], \qquad 0 \leq s \leq 1,$$

be the probability generating function of the offspring variable. Generating functions are especially well suited to branching processes because independent sums turn into composition.

Theorem 15.2. *If $G_n(s) = \mathbb{E}[s^{Z_n}]$ is the pgf of the population size in generation n, then*

$$G_{n+1}(s) = G_n(f(s)).$$

In particular, if $Z_0 = 1$, then

$$G_n(s) = f^{\circ n}(s),$$

the n-fold iterate of f.

Proof. Condition on Z_n. If $Z_n = m$, then

$$Z_{n+1} = \xi_{n,1} + \cdots + \xi_{n,m},$$

so by independence,

$$\mathbb{E}[s^{Z_{n+1}} \mid Z_n = m] = f(s)^m.$$

Therefore,

$$G_{n+1}(s) = \mathbb{E}\big[\mathbb{E}[s^{Z_{n+1}} \mid Z_n]\big] = \mathbb{E}[f(s)^{Z_n}] = G_n(f(s)).$$

If $Z_0 = 1$, then $G_0(s) = s$, so repeated composition gives the second statement. □

This theorem is one of the cleanest examples in probability of a dynamical system appearing naturally through generating functions.

15.4 Mean behavior

Let

$$m = \mathbb{E}[\xi]$$

be the mean number of offspring per individual. This single number governs much of the large-scale behavior.

Theorem 15.3. *If $\mathbb{E}[\xi] = m < \infty$, then*

$$\mathbb{E}[Z_n \mid Z_0] = Z_0 m^n.$$

In particular, if $Z_0 = 1$, then

$$\mathbb{E}[Z_n] = m^n.$$

Proof. Condition on Z_n:

$$\mathbb{E}[Z_{n+1} \mid Z_n] = \mathbb{E}\left[\sum_{k=1}^{Z_n} \xi_{n,k} \middle| Z_n\right] = Z_n \mathbb{E}[\xi] = m Z_n.$$

Taking expectations gives

$$\mathbb{E}[Z_{n+1}] = m\mathbb{E}[Z_n].$$

Iterating yields the formula. □

The cases $m < 1$, $m = 1$, and $m > 1$ are called *subcritical*, *critical*, and *supercritical*, respectively.

- If $m < 1$, the expected population decays geometrically.
- If $m = 1$, the expected population is constant.
- If $m > 1$, the expected population grows exponentially.

Yet mean growth alone does not determine survival. Even in the supercritical case, extinction may still occur with positive probability.

15.5 Extinction probability

Let

$$q = \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1).$$

The extinction event occurs if and only if every child of the initial ancestor produces a line that eventually dies out. This leads to a fixed-point equation.

Theorem 15.4. *The extinction probability q is the smallest solution in $[0,1]$ of*

$$q = f(q).$$

Proof. Condition on the number of children ξ of the initial ancestor. If the ancestor has k children, then extinction occurs exactly when each of those k descendant lines becomes extinct. By independence, that conditional probability is q^k. Therefore

$$q = \sum_{k=0}^{\infty} \mathbb{P}(\xi = k) q^k = f(q).$$

The smallest solution interpretation follows by iterating f from 0, since the probability of extinction by generation n is $f^{\circ n}(0)$. □

This fixed-point description is powerful and geometrically intuitive.

15.6 The critical trichotomy

The extinction behavior is determined by the mean offspring number.

Theorem 15.5. *Assume $\mathbb{P}(\xi = 1) \neq 1$.*

(i) If $m \leq 1$, then $q = 1$.

(ii) If $m > 1$, then $q < 1$.

Idea of proof. The pgf f is increasing and convex on $[0,1]$, with $f(1) = 1$. If $m = f'(1) \leq 1$, the graph of f lies above its tangent structure in such a way that the only fixed point in $[0,1]$ is 1. If $m > 1$, the slope at 1 is greater than 1, so the convex graph must cross the diagonal at a smaller point $q < 1$ as well. □

Thus:

- subcritical and critical processes die out almost surely;
- supercritical processes survive forever with positive probability.

The subtlety is that even when the expected population grows, randomness can still wipe out the process early.

15.7 Examples of extinction calculations

15.7.1 Binary splitting or death

Suppose an individual has either 0 children or 2 children, with probabilities $1-p$ and p. Then

$$f(s) = (1-p) + ps^2.$$

The extinction probability solves

$$q = (1-p) + pq^2.$$

Rearranging,

$$pq^2 - q + (1-p) = 0.$$

The two roots are 1 and $(1-p)/p$. The smallest root in $[0,1]$ is therefore

$$q = \begin{cases} 1, & p \le 1/2, \\ \dfrac{1-p}{p}, & p > 1/2. \end{cases}$$

The mean offspring number is $m = 2p$, so this agrees perfectly with the critical threshold $m = 1$.

15.7.2 Poisson offspring

If $\xi \sim \mathrm{Pois}(\lambda)$, then

$$f(s) = e^{\lambda(s-1)}.$$

So the extinction probability satisfies

$$q = e^{\lambda(q-1)}.$$

This equation usually cannot be solved in closed form, but its qualitative behavior is clear:

- if $\lambda \le 1$, then $q = 1$;
- if $\lambda > 1$, then there is a unique smaller solution $q < 1$.

15.8 A normalized branching-process martingale

Branching processes lead naturally to martingales. Assume $m = \mathbb{E}[\xi] \in (0, \infty)$ and define

$$W_n = \frac{Z_n}{m^n}.$$

Then

$$\mathbb{E}[W_{n+1} \mid Z_n] = \frac{1}{m^{n+1}} \mathbb{E}[Z_{n+1} \mid Z_n] = \frac{mZ_n}{m^{n+1}} = \frac{Z_n}{m^n} = W_n.$$

Thus $\{W_n\}$ is a martingale.

This observation is simple, but it is profound. It says that after scaling out the expected exponential growth or decay, the remaining process has the fair-game property.

To appreciate this fully, we now step back and discuss martingales more systematically.

15.9 What is a martingale?

A martingale is a stochastic process whose conditional expected future value equals its present value.

Definition 15.6. Let $\{\mathcal{F}_n\}$ be an increasing sequence of σ-fields, representing information revealed over time. A sequence of integrable random variables $\{M_n\}$ is a *martingale* with respect to $\{\mathcal{F}_n\}$ if:

(i) M_n is $\mathcal{F}_n$-measurable for each n;

(ii) $\mathbb{E}|M_n| < \infty$ for each n;

(iii)
$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n \qquad \text{almost surely}$$
for each n.

If the conditional expectation is at least M_n, the process is a *submartingale*; if it is at most M_n, it is a *supermartingale*.

15.10 Interpretation as a fair game

Imagine a gambler whose fortune after n rounds is M_n, and let $\mathcal{F}_n$ be the information available after round n. Then the martingale property says: given everything known so far, the expected fortune after one more round is exactly the current fortune.

This is why martingales are often described as fair games. One must be careful, however: fairness in this technical sense does not mean the process is unchanging. It may fluctuate wildly. The property is about conditional expectation, not about certainty.

15.11 Basic examples

15.11.1 Centered partial sums

Let $X_1, X_2, \dots$ be independent integrable random variables with mean 0, and define
$$M_n = X_1 + \cdots + X_n, \qquad \mathcal{F}_n = \sigma(X_1, \dots, X_n).$$
Then
$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n + \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = M_n + \mathbb{E}[X_{n+1}] = M_n.$$

So the centered partial sums form a martingale.

In particular, simple symmetric random walk is a martingale.

15.11.2 Conditional expectations of a fixed variable

Let Y be integrable and define

$$M_n = \mathbb{E}[Y \mid \mathcal{F}_n].$$

Then $\{M_n\}$ is a martingale. This follows immediately from the tower property:

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[Y \mid \mathcal{F}_n] = M_n.$$

This is the conceptual prototype of martingales. As information increases, the best prediction of a fixed integrable quantity evolves as a martingale.

15.11.3 Branching-process normalization

As noted above,

$$W_n = Z_n / m^n$$

is a martingale for a Galton–Watson process with mean offspring number m.

15.12 Submartingales and convex functions

If M_n is a martingale and φ is convex, then under suitable integrability conditions, $\varphi(M_n)$ is a submartingale. The reason is conditional Jensen's inequality:

$$\mathbb{E}[\varphi(M_{n+1}) \mid \mathcal{F}_n] \geq \varphi(\mathbb{E}[M_{n+1} \mid \mathcal{F}_n]) = \varphi(M_n).$$

A very useful special case is that if M_n is a martingale, then $|M_n|$ and M_n^2 are submartingales whenever integrable.

This simple fact leads to many inequalities and convergence theorems in more advanced courses.

15.13 Optional stopping: the basic idea

One of the most attractive martingale principles is that a fair game remains fair even if we choose a random time to stop, provided that time is sufficiently well behaved.

Definition 15.7. A random time T taking values in $\{0, 1, 2, \dots\} \cup \{\infty\}$ is a *stopping time* with respect to $\{\mathcal{F}_n\}$ if for each n the event $\{T = n\}$ depends only on information available by time n.

Equivalently, the event $\{T \leq n\}$ must belong to $\mathcal{F}_n$ for each n.

Examples include:

- the first time a random walk hits a specified state;
- the first time a process exceeds a threshold;
- a deterministic time n.

Not every random time is a stopping time. "The last time the process visits 0" typically is not, because knowing that a time is the last visit requires future information.

15.14 A bounded optional stopping theorem

The full optional stopping theorem has several versions. The cleanest undergraduate form is the bounded case.

Theorem 15.8 (Optional stopping, bounded case)**.** *Let $\{M_n\}$ be a martingale with respect to $\{\mathcal{F}_n\}$, and let T be a stopping time bounded by some constant N. Then*

$$\mathbb{E}[M_T] = \mathbb{E}[M_0].$$

Idea of proof. One considers the stopped process

$$M_{n\wedge T} = M_{\min(n,T)}.$$

This is again a martingale. Since $T \le N$, we have $M_T = M_{N\wedge T}$, so

$$\mathbb{E}[M_T] = \mathbb{E}[M_{N\wedge T}] = \mathbb{E}[M_0].$$

The essential point is that stopping at a bounded time does not destroy the fair-game property. □

This theorem is both intuitive and powerful. It converts martingale structure into concrete formulas.

15.15 Application to simple random walk

Let $S_n = X_1 + \cdots + X_n$ be simple symmetric random walk, where the X_i are i.i.d. with

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}.$$

Then $\{S_n\}$ is a martingale.

Suppose we start at $S_0 = i$ and stop when the walk first hits either 0 or N. Let

$$T = \inf\{n \ge 0 : S_n \in \{0, N\}\}.$$

Formally T need not be bounded, so the bounded optional stopping theorem does not apply directly. But if one first truncates by $T \wedge m$ and then lets $m \to \infty$, one can justify the following classical result.

Theorem 15.9 (Fair random walk hitting probability)**.** *For simple symmetric random walk on $\{0, 1, \dots, N\}$ absorbed at the endpoints,*

$$\mathbb{P}_i(\tau_N < \tau_0) = \frac{i}{N}.$$

Martingale proof sketch. Since S_n is a martingale,

$$\mathbb{E}_i[S_{T \wedge m}] = i.$$

As $m \to \infty$, one obtains $\mathbb{E}_i[S_T] = i$. But S_T equals either 0 or N, so if $h(i) = \mathbb{P}_i(\tau_N < \tau_0)$,

$$\mathbb{E}_i[S_T] = 0 \cdot (1 - h(i)) + N \cdot h(i) = Nh(i).$$

Thus $Nh(i) = i$, giving $h(i) = i/N$. □

This proof is short and elegant. It shows how martingales can replace difference-equation calculations.

15.16 A second martingale for random walk

For simple symmetric random walk, the process

$$M_n = S_n^2 - n$$

is also a martingale. Indeed,

$$S_{n+1} = S_n + X_{n+1},$$

so

$$S_{n+1}^2 = S_n^2 + 2S_nX_{n+1} + X_{n+1}^2.$$

Taking conditional expectation given $\mathcal{F}_n$ and using $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = 0$ and $X_{n+1}^2 = 1$ gives

$$\mathbb{E}[S_{n+1}^2 - (n+1) \mid \mathcal{F}_n] = S_n^2 - n.$$

This martingale yields expected hitting-time formulas. For instance, in gambler's ruin with fair steps,

$$\mathbb{E}_i[T] = i(N - i)$$

can be derived from optional stopping applied to $S_n^2 - n$.

15.17 Martingales from Markov chains

Suppose X_n is a Markov chain and f is a function on the state space. If

$$(Pf)(i) = \sum_j p_{ij} f(j) = f(i)$$

for all states i, then

$$\mathbb{E}[f(X_{n+1}) \mid X_n] = f(X_n),$$

so $f(X_n)$ is a martingale.

Such functions are called *harmonic* for the chain. They arise naturally in hitting-probability problems. For example, in gambler's ruin, the function $h(i) = i/N$ is harmonic for fair random walk on the interior states.

This is another way to understand first-step analysis: solving a hitting-probability problem often amounts to finding a bounded harmonic function with given boundary values.

15.18 Why martingales matter beyond this course

Martingales appear throughout advanced probability because they provide a general framework for sequential conditioning. They underlie:

- convergence theorems;
- concentration inequalities;
- stopping-time arguments;
- stochastic integration and Itô calculus;
- modern statistical and algorithmic analyses.

The formal theory is much deeper than what we present here, but the basic examples already show the flavor: when a process has the right conditional expectation structure, powerful conclusions often follow.

15.19 Summary

This final chapter offered two glimpses beyond the main syllabus.

- A Galton–Watson branching process evolves by independent offspring reproduction.
- Its pgf evolves by composition, and the extinction probability is the smallest fixed point of the offspring pgf.

- The mean offspring number determines whether extinction occurs almost surely or survival is possible with positive probability.
- A martingale is a process whose conditional expected next value equals its present value.
- Centered sums, conditional expectations, and normalized branching processes are basic martingale examples.
- Optional stopping and martingale constructions give elegant solutions to classical random-walk problems.

A strong undergraduate course need not treat every later topic in full detail. But students should leave recognizing that conditional expectation is not the end of the story; it is the beginning of a much larger part of probability.

Exercises

Exercise 15.1. Let ξ be the offspring variable in a Galton–Watson process. Define the offspring pgf

$$f(s) = \mathbb{E}[s^{\xi}].$$

Explain why

$$\mathbb{E}[s^{Z_{n+1}} \mid Z_n] = f(s)^{Z_n}.$$

Exercise 15.2. Suppose each individual has 0 children with probability $1/3$ and 2 children with probability $2/3$. Compute the mean offspring number and determine whether extinction occurs almost surely.

Exercise 15.3. For the same offspring distribution as in the previous exercise, solve for the extinction probability.

Exercise 15.4. Show that if $\mathbb{E}[\xi] = m$, then $\mathbb{E}[Z_n] = m^n$ when $Z_0 = 1$.

Exercise 15.5. Let $X_1, X_2, \dots$ be independent with mean 0, and define $M_n = X_1 + \cdots + X_n$. Show that $\{M_n\}$ is a martingale with respect to $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$.

Exercise 15.6. Let Y be integrable and define $M_n = \mathbb{E}[Y \mid \mathcal{F}_n]$. Show directly from the tower property that $\{M_n\}$ is a martingale.

Exercise 15.7. For simple symmetric random walk S_n, verify that $S_n^2 - n$ is a martingale.

Exercise 15.8. What is a stopping time? Give two examples and one example of a random time that is not a stopping time.

Exercise 15.9. State the bounded optional stopping theorem in your own words.

Exercise 15.10. Consider fair gambler's ruin on $\{0, 1, 2, 3, 4\}$ starting from 2. Use the martingale idea to compute the probability of hitting 4 before 0.

Challenge Exercise 15.11. Suppose a branching process has Poisson(λ) offspring. Show that the extinction probability is the smallest solution of

$$q = e^{\lambda(q-1)}.$$

Explain graphically why there is a solution $q < 1$ exactly when $\lambda > 1$.

Challenge Exercise 15.12. Let X_n be a Markov chain with transition matrix P. Suppose f satisfies $Pf = f$. Show that $f(X_n)$ is a martingale. Then explain how harmonic functions are connected to hitting probabilities.

Chapter A

A Short Foundations Primer

A first probability course should not drown students in abstraction, but it should also not hide the mathematical structure that makes the subject coherent. This appendix collects the foundational ideas that sit quietly underneath the main text: σ-fields, measurability, probability measures, expectation as integration, and the logic of conditional expectation.

A.1 Why probability needs foundations

In elementary examples, probability can appear to be nothing more than counting equally likely outcomes or integrating densities. But those formulas work only after a model has been built carefully enough for the formulas to make sense.

Several questions force us to think more structurally.

- Which subsets of the sample space are we allowed to assign probabilities to?
- Why is a random variable defined by a measurability condition rather than just being any function?
- What exactly is expectation when the random variable is not discrete?
- How can conditional expectation be defined when one conditions on a random variable that takes continuum-many values?

The answers lie in measure-theoretic probability. A full course in measure theory is not required for understanding the main undergraduate narrative, but the ideas are worth seeing in a compact, readable form.

A.2 Algebras and σ-algebras

Let Ω be a sample space. A family of subsets of Ω is called an *algebra* if it contains Ω, is closed under complements, and is closed under finite unions. A family is called a *σ-algebra* if, in addition, it is closed under countable unions.

Definition A.1. A collection $\mathcal{F} \subseteq 2^{\Omega}$ is a *σ-algebra* if:

(i) $\Omega \in \mathcal{F}$;

(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;

(iii) if $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

From these axioms it follows automatically that $\varnothing \in \mathcal{F}$, that $\mathcal{F}$ is closed under countable intersections, and that it is closed under set differences.

A.2.1 Why countable operations?

Countable closure is not a technical luxury. It is exactly what is needed for limits. Events defined by convergence, limsup, liminf, and continuity are built through countable unions and intersections. If we want probability to interact properly with asymptotic reasoning, a mere algebra is not enough.

Example A.2. If X_n is a sequence of random variables and we want the event

$$\{X_n \to X\},$$

we need countable unions and intersections to describe it. One representation is

$$\bigcap_{m=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n \geq N} \{|X_n - X| < 1/m\}.$$

Thus the event that a sequence converges is naturally a σ-algebra event.

A.3 Generated σ-algebras

If $\mathcal{C}$ is a collection of subsets of Ω, the σ-algebra *generated* by $\mathcal{C}$, denoted $\sigma(\mathcal{C})$, is the smallest σ-algebra containing $\mathcal{C}$.

This construction appears constantly.

- The Borel σ-algebra on $\mathbb{R}$ is generated by intervals.
- The information revealed by a random variable X is the σ-algebra generated by the events $\{X \in B\}$.
- A filtration $\{\mathcal{F}_n\}$ is often defined by generated σ-algebras from observed data.

A.4 Probability measures

A probability measure is a countably additive way of assigning mass to events.

Definition A.3. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a *probability measure* if

$$\mathbb{P}(\Omega) = 1$$

and for every sequence of pairwise disjoint events $A_1, A_2, \cdots \in \mathcal{F}$,

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

Countable additivity is the essential axiom. Finite additivity alone would be too weak for limit theorems and for defining integration properly.

A.4.1 Basic consequences

From the axioms one obtains:

- monotonicity: if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$;
- continuity from below: if $A_n \uparrow A$, then $\mathbb{P}(A_n) \to \mathbb{P}(A)$;
- continuity from above: if $A_n \downarrow A$ and $\mathbb{P}(A_1) < \infty$, then $\mathbb{P}(A_n) \to \mathbb{P}(A)$.

These continuity properties are used constantly, even when not named.

A.5 The Borel σ-algebra

On the real line, the natural measurable sets are the Borel sets.

Definition A.4. The *Borel σ-algebra* on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the σ-algebra generated by the open intervals.

One may equivalently generate it using closed intervals, half-lines of the form $(-\infty, a]$, or open sets. The equivalence matters because it makes measurability checks easy.

Proposition A.5. *A real-valued function X on $(\Omega, \mathcal{F})$ is measurable if and only if for every real a,*

$$\{X \leq a\} \in \mathcal{F}.$$

This criterion is one of the most useful little theorems in the subject.

A.6 Random variables as measurable functions

A random variable is not merely a function from outcomes to numbers. It must be compatible with the event structure.

Definition A.6. A function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a *random variable* if for every Borel set B,

$$\{X \in B\} = X^{-1}(B) \in \mathcal{F}.$$

The point of measurability is that probabilities such as $\mathbb{P}(X \in B)$ make sense only if the inverse image $X^{-1}(B)$ is an event.

A.6.1 Why this definition is natural

Suppose we want to discuss the cdf of X:

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

This quantity makes sense only if the set $\{X \leq x\}$ lies in the event σ-field. Measurability guarantees exactly that.

A.6.2 Closure properties

If X and Y are measurable, then so are $X+Y$, XY, $\max\{X, Y\}$, $\min\{X, Y\}$, and pointwise limits of measurable functions. These closure properties make random variables stable under the operations one naturally wants to perform.

A.7 Distribution measures

Every random variable induces a probability measure on the real line.

Definition A.7. The *distribution* or *law* of a random variable X is the probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$\mu_X(B) = \mathbb{P}(X \in B).$$

This point of view is conceptually useful. It separates:

- the underlying experiment $(\Omega, \mathcal{F}, \mathbb{P})$;
- the induced law of the numerical quantity X.

Many theorems depend only on the law of X, not on the particular sample space on which X is realized.

A.8 Expectation as an integral

For discrete random variables one first meets expectation as a weighted average. In full generality, expectation is integration with respect to a probability measure.

The modern construction proceeds in stages.

A.8.1 Step 1: simple functions

A *simple function* is a measurable function of the form

$$Y = \sum_{k=1}^{m} a_k \mathbf{1}_{A_k}$$

with measurable sets A_k and real coefficients a_k. For nonnegative simple functions define

$$\mathbb{E}[Y] = \sum_{k=1}^{m} a_k \mathbb{P}(A_k).$$

This agrees with the intuitive discrete formula.

A.8.2 Step 2: nonnegative measurable functions

If $X \geq 0$ is measurable, define

$$\mathbb{E}[X] = \sup\{\mathbb{E}[Y] : 0 \leq Y \leq X,\ Y \text{ simple}\}.$$

This is the Lebesgue integral of X with respect to $\mathbb{P}$.

A.8.3 Step 3: integrable real-valued functions

For arbitrary X, write

$$X = X^+ - X^-, \qquad X^+ = \max\{X, 0\}, \quad X^- = -\min\{X, 0\}.$$

If both $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are finite, then X is integrable and

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

This construction may look formal, but it makes expectation work equally well for discrete, continuous, and mixed distributions.

A.9 Why Lebesgue integration is better suited to probability

The Riemann integral is excellent for smooth deterministic calculus, but the Lebesgue integral is better adapted to probability because:

- it handles indicator functions and discontinuities naturally;
- it interacts cleanly with limits of random variables;
- it is built directly from measure, so probabilities and expectations live in the same framework.

Probability is full of limit operations, and the main convergence theorems of integration are therefore indispensable.

A.10 Monotone and dominated convergence

Two theorems deserve special mention.

Theorem A.8 (Monotone convergence)**.** *If $0 \le X_n \uparrow X$ pointwise, then*

$$\mathbb{E}[X_n] \uparrow \mathbb{E}[X].$$

Theorem A.9 (Dominated convergence)**.** *If $X_n \to X$ almost surely and there exists an integrable random variable Y such that $|X_n| \le Y$ for all n, then*

$$\mathbb{E}[X_n] \to \mathbb{E}[X].$$

These theorems are the machinery behind many seemingly elementary probability arguments. For instance, the proof that almost sure convergence implies convergence in probability by indicator functions is often followed by dominated convergence. The derivation of moment formulas from mgfs or characteristic functions also rests on justified limit exchanges.

A.11 Independence as product structure

Independence is often introduced by formulas such as

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

At a deeper level, independence means product structure.

If X and Y are random variables, they are independent if and only if their joint law factors as the product of their marginal laws:

$$\mu_{(X,Y)} = \mu_X \otimes \mu_Y.$$

Equivalently,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all Borel sets A, B.

This product viewpoint explains why expectations factor under independence:

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

whenever the expectations are well defined.

A.12 Product spaces

To build independent random variables abstractly, one often uses product spaces. If $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ are probability spaces, their product space is

$$(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2).$$

The coordinate maps are then independent random elements. This formalism underlies the construction of sequences of independent trials.

A.13 Conditional expectation as an abstract object

In simple discrete settings one can define

$$\mathbb{E}[X \mid Y = y]$$

by conditioning on the event $\{Y = y\}$ and weighting over the conditional distribution. That works well when Y is discrete and $\mathbb{P}(Y = y) > 0$. In general, however, many conditioning events have probability zero. For example, if Y has a density, then $\mathbb{P}(Y = y) = 0$ for every y.

The correct general object is conditional expectation with respect to a σ-field.

Definition A.10. Let X be integrable and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-field. A random variable Z is called a version of $\mathbb{E}[X \mid \mathcal{G}]$ if:

(i) Z is $\mathcal{G}$-measurable;

(ii) for every $G \in \mathcal{G}$,

$$\mathbb{E}[X\mathbf{1}_G] = \mathbb{E}[Z\mathbf{1}_G].$$

The second condition says that Z reproduces the integrals of X on every event whose truth is determined by the information in $\mathcal{G}$.

A.13.1 Uniqueness up to almost sure equality

Conditional expectation is unique only almost surely. If Z and Z' both satisfy the defining properties, then

$$Z = Z' \qquad \text{almost surely.}$$

This is not a defect. In probability, random variables that differ only on a null set are usually regarded as representing the same mathematical quantity.

A.14 Conditional expectation given a random variable

If Y is a random variable, then the information revealed by observing Y is encoded by the generated σ-field

$$\sigma(Y) = \{Y^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}.$$

Then

$$\mathbb{E}[X \mid Y]$$

really means

$$\mathbb{E}[X \mid \sigma(Y)].$$

A basic theorem says that any $\sigma(Y)$-measurable random variable can be written as a measurable function of Y. Therefore there exists a Borel function g such that

$$\mathbb{E}[X \mid Y] = g(Y) \qquad \text{almost surely.}$$

This justifies the notation used throughout the main text.

A.15 The tower property from the measure-theoretic viewpoint

If $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$, then

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}].$$

This is the tower property in full generality. It follows directly from the defining integral property of conditional expectation. In words: if one first conditions on more information and then forgets part of it, the result is the same as conditioning directly on the smaller information set.

A.16 Conditional expectation as projection in L^2

When $X \in L^2$, there is a geometric interpretation. The collection of square-integrable $\mathcal{G}$-measurable random variables forms a closed subspace of the Hilbert space L^2. Then $\mathbb{E}[X \mid \mathcal{G}]$ is the orthogonal projection of X onto that subspace.

This explains why conditional expectation is the best mean-square predictor given the information

in $\mathcal{G}$. It also explains identities such as the law of total variance and conditional Jensen in a conceptually satisfying way.

A.17 Null sets and almost sure statements

Probability theory is full of statements that hold *almost surely*, meaning outside a set of probability zero. This language is essential because measure-theoretic constructions naturally identify functions up to null sets.

For example:

- two versions of conditional expectation may differ on a null set;
- convergence almost surely ignores exceptional outcomes of probability zero;
- densities are unchanged by modifying them on sets of Lebesgue measure zero.

One should not think of null sets as irrelevant in every context, but in most of probability they are mathematically negligible.

A.18 Why uncountable sample spaces are subtle

On a finite or countable sample space one can often assign probabilities to all subsets. On uncountable spaces this is impossible in many natural settings if one wants countable additivity and geometric consistency. That is why we work with carefully chosen σ-algebras such as the Borel sets, rather than with all subsets of $\mathbb{R}$.

This fact may feel surprising at first, but it is one of the reasons measure theory is not optional in rigorous probability.

A.19 How much of this appendix is needed?

For most first-course computations, not much. But conceptually, a lot.

A student who remembers the following points already has a very strong foundation:

(1) Events live in a σ-algebra because countable operations matter.

(2) Random variables are measurable functions so that inverse images of Borel sets are events.

(3) Expectation is integration with respect to a probability measure.

(4) Independence means product structure.

(5) Conditional expectation is defined relative to information, encoded by a sub-σ-field.

Those five ideas quietly support nearly everything in the main text.

A.20 Summary

This appendix provided a compact map of the rigorous structure behind undergraduate probability.

- σ-algebras organize events and make limit events measurable.
- Probability measures are countably additive set functions.
- Random variables are measurable maps into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- Expectation is Lebesgue integration.
- Convergence theorems justify many limit interchanges.
- Independence is product structure.
- Conditional expectation is best understood relative to a sub-σ-field.

These ideas are not separate from elementary probability. They are the clean mathematical language in which elementary probability becomes fully coherent.

Chapter B

Common Distributions, Identities, and Problem-Solving Tools

A large part of learning probability is learning which structures recur. This appendix collects the families, formulas, and strategic moves that appear again and again. It is intended not as a substitute for understanding, but as a compact reference to support problem solving and review.

B.1 A note on using formula sheets wisely

Students often ask for a formula sheet. Such a sheet is useful, but only if it organizes the subject rather than replacing it. The right question is not "What formula matches this problem?" but "What structure does this problem have?"

The formulas below are therefore grouped by conceptual role.

- distributions and their basic properties;
- standard analytic identities;
- recurring proof methods;
- approximation checklists.

B.2 Core notation

Notation	Meaning
Ω	sample space
$\mathcal{F}$	event σ-algebra

$\mathbb{P}$	probability measure
X, Y, Z	random variables
F_X	cdf of X, $F_X(x) = \mathbb{P}(X \le x)$
p_X	pmf of a discrete random variable
f_X	pdf of a continuous random variable
$\mathbb{E}[X]$	expectation of X
$\mathrm{Var}(X)$	variance of X
$\mathrm{Cov}(X, Y)$	covariance of X and Y
$\mathrm{Corr}(X, Y)$	correlation of X and Y
$\mathbf{1}_A$	indicator of event A
Φ	standard normal cdf
ϕ	standard normal pdf
$\xrightarrow{\mathbb{P}}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
$\xrightarrow{a.s.}$	almost sure convergence
$\mathrm{Bin}(n, p)$	binomial distribution
$\mathrm{Pois}(\lambda)$	Poisson distribution
$\mathrm{Exp}(\lambda)$	exponential distribution with rate λ
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2

B.3 Common discrete distributions

Name	pmf	Support	Mean	Variance	Comments
Bernoulli(p)	$\mathbb{P}(X=1)=p$, $\mathbb{P}(X=0)=1-p$	$\{0,1\}$	p	$p(1-p)$	indicator model
Binomial(n,p)	$\binom{n}{k}p^k(1-p)^{n-k}$	$k=0,\dots,n$	np	$np(1-p)$	sum of n i.i.d. Bernoullis
Geometric(p)	$(1-p)^{k-1}p$	$k=1,2,\dots$	$1/p$	$(1-p)/p^2$	waiting time to first success
Negative Binomial(r,p)	$\binom{k-1}{r-1}p^r(1-p)^{k-r}$	$k=r,r+1,\dots$	r/p	$r(1-p)/p^2$	waiting time to rth success
Hypergeometric (N,M,n)	$\dfrac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}$	feasible k	$n\dfrac{N}{N+M}$	$n\dfrac{N}{N+M}\times\dfrac{M}{N+M}\times\dfrac{N+M-n}{N+M-1}$	sampling without replacement
Poisson(λ)	$e^{-\lambda}\dfrac{\lambda^k}{k!}$	$k=0,1,2,\dots$	λ	λ	rare-event count
Discrete uniform on $\{1,\dots,n\}$	$1/n$	$1,\dots,n$	$(n+1)/2$	$(n^2-1)/12$	symmetric finite model

B.4 Common continuous distributions

Name	Density	Support	Mean	Variance	Comments
Uniform(a,b)	$\dfrac{1}{b-a}$	$a<x<b$	$(a+b)/2$	$(b-a)^2/12$	flat density
Exponential(λ)	$\lambda e^{-\lambda x}$	$x>0$	$1/\lambda$	$1/\lambda^2$	memoryless
Normal(μ,σ^2)	$\dfrac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$	$x\in\mathbb{R}$	μ	σ^2	CLT limit law
Gamma(α,λ)	$\dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$x>0$	α/λ	α/λ^2	sum of exponentials when α is an integer
Beta(α,β)	$\dfrac{1}{B(\alpha,\beta)}$ $\times x^{\alpha-1}(1-x)^{\beta-1}$	$0<x<1$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2}$ $\times\dfrac{1}{\alpha+\beta+1}$	flexible distribution on $(0,1)$
Chi-square(ν)	$\dfrac{1}{2^{\nu/2}\Gamma(\nu/2)}$ $\times x^{\nu/2-1}e^{-x/2}$	$x>0$	ν	2ν	Gamma special case

B.5 Generating functions and transforms at a glance

Object	Definition and main use
Probability generating function	For nonnegative integer-valued X, $G_X(s)=\mathbb{E}[s^X]$. Useful for sums of independent count variables, branching processes, and obtaining moments via derivatives at $s=1$.
Moment generating function	$M_X(t)=\mathbb{E}[e^{tX}]$ when finite near 0. Useful for moments, for identifying distributions, and for proving the CLT under an extra regularity assumption.
Characteristic function	$\varphi_X(t)=\mathbb{E}[e^{itX}]$. Always exists. Useful for distribution identification, convolution, and general limit theorems.

B.5.1 Selected transform formulas

Distribution	Useful transform
Bernoulli(p)	$G(s)=1-p+ps,\quad M(t)=1-p+pe^t$
Binomial(n,p)	$G(s)=(1-p+ps)^n,\quad M(t)=(1-p+pe^t)^n$

Geometric(p)	$G(s) = \dfrac{ps}{1-(1-p)s}$ for $\lvert s\rvert < 1/(1-p)$
Poisson(λ)	$G(s) = e^{\lambda(s-1)}, \quad M(t) = e^{\lambda(e^t-1)}, \quad \varphi(t) = e^{\lambda(e^{it}-1)}$
Exponential(λ)	$M(t) = \dfrac{\lambda}{\lambda - t}$ for $t < \lambda, \quad \varphi(t) = \dfrac{\lambda}{\lambda - it}$
Normal(μ, σ^2)	$M(t) = e^{\mu t + \sigma^2 t^2/2}, \quad \varphi(t) = e^{i\mu t - \sigma^2 t^2/2}$
Gamma(α, λ)	$M(t) = \left(\dfrac{\lambda}{\lambda - t}\right)^{\alpha}$ for $t < \lambda$

B.6 Useful series and analytic identities

Identity	Use
$e^x = \sum_{n=0}^{\infty} \dfrac{x^n}{n!}$	foundational in Poisson, mgf, and characteristic-function work
$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$	binomial expansions, combinatorics
$\sum_{k=0}^{\infty} r^k = \dfrac{1}{1-r}$ for $\lvert r\rvert < 1$	geometric distributions, pgf algebra
$\sum_{k=1}^{\infty} k r^{k-1} = \dfrac{1}{(1-r)^2}$	moments of geometric-type laws
$\log(1+x) = x + o(x)$ as $x \to 0$	limit proofs for Poisson approximation and CLT transforms
$(1 + \frac{x}{n})^n \to e^x$	rare-event and exponential limits
$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$	Gamma calculations
$B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$	Beta normalization and moments

B.7 Stirling's approximation

For large n,

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

This approximation is extremely useful for asymptotic analysis of factorial expressions, especially in binomial coefficients and local limit estimates. In an undergraduate course it often appears heuristically rather than as a proved theorem.

B.8 Expectation identities worth memorizing

(1) **Linearity:**

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c.$$

(2) **Indicator trick:**

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

(3) **Expectation of a sum of indicators:**

$$\mathbb{E}\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} \mathbb{P}(I_i = 1).$$

This is often the fastest way to compute expected counts.

(4) **Law of total expectation:**

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

(5) **Tail-sum formula for nonnegative integer-valued X:**

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

(6) **Variance formula:**

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

(7) **Variance of independent sums:**

$$\operatorname{Var}\left(\sum_i X_i\right) = \sum_i \operatorname{Var}(X_i)$$

when the X_i are independent.

(8) **Law of total variance:**

$$\operatorname{Var}(X) = \mathbb{E}[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(\mathbb{E}[X \mid Y]).$$

B.9 Standard inequalities

Inequality	Statement and use
Union bound	$\mathbb{P}(\bigcup_i A_i) \leq \sum_i \mathbb{P}(A_i)$. Basic but powerful.
Markov	If $X \geq 0$, then $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$. Turns mean bounds into tail bounds.
Chebyshev	$\mathbb{P}(\lvert X - \mu \rvert \geq a) \leq \operatorname{Var}(X)/a^2$. Main tool for weak laws.
Cauchy–Schwarz	$\lvert \mathbb{E}[XY] \rvert \leq (\mathbb{E}[X^2]\mathbb{E}[Y^2])^{1/2}$. Controls covariance and integrability.
Jensen	For convex φ, $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$. Useful for inequalities and conditioning.

Hoeffding (bounded case)	Gives exponential concentration for bounded independent sums. Useful when Chebyshev is too crude.

B.10 Recognizing standard structures in problems

A large share of probability problem solving is pattern recognition. The following cues are worth practicing.

B.10.1 Indicator structure

If a problem asks for the expected number of objects having some property, define indicators and sum them. This avoids complicated dependence.

Typical phrases:

- expected number of matches;
- expected number of empty boxes;
- expected number of fixed points;
- expected number of runs.

B.10.2 Conditioning structure

If a problem contains partial information, multiple stages, or a difficult distribution that becomes simple after conditioning, use the law of total probability or the law of total expectation.

Typical phrases:

- first choose a box, then draw a ball;
- given the total number of trials;
- conditional on the first step;
- mixture model.

B.10.3 Symmetry structure

If outcomes are exchangeable or symmetric, the answer is often determined without heavy computation.

Typical examples:

- each permutation position is equally likely;

- each player is equally likely to win under fair rules;
- random walk started at the midpoint of symmetric boundaries.

B.10.4 Transform structure

If the question involves sums of independent random variables, generating functions, mgfs, or characteristic functions may linearize the problem.

B.10.5 First-step structure

If a stochastic process evolves recursively and the question concerns hitting or absorption, condition on the first step.

B.11 How to choose among common methods

Problem type	Good first method
Expected count of events	indicator variables
Distribution of a sum of independent discrete counts	convolution or pgf
Waiting time until first success	geometric or exponential model
Hitting probability in a chain or random walk	first-step analysis or martingale
Approximate count of rare events	Poisson approximation
Approximate average of many i.i.d. variables	central limit theorem
Long-run distribution of a finite irreducible aperiodic chain	stationary distribution + convergence theorem
Conditional average after observing information	conditional expectation / tower property

B.12 Approximation checklist

B.12.1 Poisson approximation

Use when:

- you have many trials or opportunities;
- each success event is individually rare;

- the total expected count is moderate.

Typical replacement:

$$\text{Bin}(n,p) \approx \text{Pois}(np) \qquad \text{when } n \text{ large and } p \text{ small.}$$

B.12.2 Normal approximation

Use when:

- you are summing many independent contributions;
- the variance is finite;
- no single term dominates the sum;
- for binomial data, both np and $n(1-p)$ are reasonably large.

Typical replacements:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx \mathcal{N}(0,1), \qquad \overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

B.12.3 Continuity correction

When approximating a discrete law by a continuous normal law, shift integer thresholds by ± 0.5.

Examples:

$$\mathbb{P}(X \le k) \approx \Phi\left(\frac{k+0.5-\mu}{\sigma}\right),$$

$$\mathbb{P}(a \le X \le b) \approx \Phi\left(\frac{b+0.5-\mu}{\sigma}\right) - \Phi\left(\frac{a-0.5-\mu}{\sigma}\right).$$

B.13 Common distributional facts that explain models

(1) Sums of independent Poisson variables are Poisson.

(2) Sums of independent normals are normal.

(3) Sums of independent exponentials of the same rate are Gamma.

(4) Conditional on a fixed total, multinomial counts arise naturally.

(5) Conditional on a Poisson count in an interval, Poisson arrival times are uniform order statistics.

Students often remember these as formulas, but they are more than formulas: they are model signatures.

B.14 A minimal checklist before finalizing a solution

When solving a probability problem, it helps to pause before writing the final answer and check the following.

(1) Did I define the relevant random variable or event clearly?

(2) If I used independence, is it actually justified?

(3) If I conditioned, did I average over all conditioning values correctly?

(4) If I used an approximation, did I state why it is appropriate?

(5) Does the final answer have the right units, scale, and range?

Many mistakes disappear at this stage.

B.15 Summary

This appendix gathered high-frequency facts and methods.

- The most important distributions have recognizable supports, parameters, means, and variances.
- Generating functions and transforms encode convolution and moments.
- A small set of series expansions and inequalities powers a large fraction of the course.
- Most problems become easier once one recognizes whether the right move is counting, conditioning, symmetry, transforms, or first-step analysis.

Use this appendix as a reference, but try to connect each formula back to the chapter where it was motivated. That is how formulas become understanding rather than clutter.

Chapter C

Cumulative Review Problems with Solution Sketches

The best way to learn probability is to solve problems, compare methods, and reflect on why each method works. This appendix provides cumulative review sets across the course. The solution sketches are intentionally compact: they are meant to guide your reasoning, not replace it.

C.1 How to use this appendix

These problems are arranged in three groups.

(1) Foundations, counting, distributions, and expectation.

(2) Conditioning, transforms, convergence, and asymptotics.

(3) Stochastic processes: Poisson processes, Markov chains, branching, and martingales.

A useful study routine is:

- first solve a problem without looking at the sketch;
- then compare your argument to the sketch;
- finally ask whether there was a cleaner route.

Probability is a subject in which elegance often comes from selecting the right viewpoint.

C.2 Part I: Foundations, counting, and expectation

Problem 1

A fair coin is tossed four times. What is the probability of exactly two heads? What is the probability of at least two heads?

Sketch. The sample space has size $2^4 = 16$. Exactly two heads can occur in

$$\binom{4}{2} = 6$$

ways, so the probability is $6/16 = 3/8$. For at least two heads, add the counts for two, three, and four heads:

$$\frac{\binom{4}{2} + \binom{4}{3} + \binom{4}{4}}{16} = \frac{6 + 4 + 1}{16} = \frac{11}{16}.$$

The important structural point is that the order matters in the sample space but not in the counting of head positions.

Problem 2

A card is drawn uniformly from a standard deck. Let A be the event "the card is a heart" and B the event "the card is a face card." Compute $\mathbb{P}(A \mid B)$.

Sketch. There are 12 face cards in the deck and 3 of them are hearts (jack, queen, king of hearts). Therefore

$$\mathbb{P}(A \mid B) = \frac{3}{12} = \frac{1}{4}.$$

This is a direct conditioning problem on a finite uniform sample space.

Problem 3

A diagnostic test is positive with probability 0.95 when a person has a disease and with probability 0.08 when the person does not have the disease. Suppose 2% of the population has the disease. What is the probability a person has the disease given a positive test?

Sketch. Let D be disease and $+$ be positive test. Bayes' rule gives

$$\mathbb{P}(D \mid +) = \frac{\mathbb{P}(+ \mid D)\mathbb{P}(D)}{\mathbb{P}(+ \mid D)\mathbb{P}(D) + \mathbb{P}(+ \mid D^c)\mathbb{P}(D^c)}.$$

So

$$\mathbb{P}(D \mid +) = \frac{0.95 \cdot 0.02}{0.95 \cdot 0.02 + 0.08 \cdot 0.98} \approx 0.195.$$

Even with a good test, rare diseases can produce modest posterior probabilities because false positives come from a much larger base population.

Problem 4

Let $X \sim \mathrm{Bin}(8, 0.3)$. Compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and $\mathbb{P}(X = 0)$.

Sketch. For a binomial random variable,

$$\mathbb{E}[X] = np = 8(0.3) = 2.4, \qquad \mathrm{Var}(X) = np(1-p) = 8(0.3)(0.7) = 1.68.$$

Also,

$$\mathbb{P}(X = 0) = (1 - 0.3)^8 = 0.7^8.$$

This is a pure distribution-recognition problem.

Problem 5

Let X be geometric with success probability p. Show directly from the pmf that $\mathbb{E}[X] = 1/p$.

Sketch. Use the pmf

$$\mathbb{P}(X = k) = (1-p)^{k-1}p, \qquad k \geq 1.$$

Then

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

With $r = 1 - p$, use the identity

$$\sum_{k=1}^{\infty} kr^{k-1} = \frac{1}{(1-r)^2}$$

for $|r| < 1$. This gives

$$\mathbb{E}[X] = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

The main lesson is that geometric-series derivatives are a recurring algebraic tool in discrete probability.

Problem 6

Suppose X has density $f(x) = 2x$ on $0 < x < 1$ and 0 otherwise. Find the cdf, the mean, and the variance.

Sketch. For $0 < x < 1$,

$$F(x) = \int_0^x 2t\,dt = x^2.$$

Also,

$$\mathbb{E}[X] = \int_0^1 x(2x)\,dx = 2\int_0^1 x^2\,dx = \frac{2}{3},$$

and

$$\mathbb{E}[X^2] = \int_0^1 x^2(2x)\,dx = 2\int_0^1 x^3\,dx = \frac{1}{2}.$$

Hence

$$\mathrm{Var}(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

Problem 7

A box contains 5 red balls and 7 blue balls. Three balls are drawn without replacement. Let X be the number of red balls drawn. Identify the distribution of X and compute $\mathbb{P}(X = 2)$.

Sketch. This is hypergeometric sampling without replacement. Thus

$$\mathbb{P}(X = 2) = \frac{\binom{5}{2}\binom{7}{1}}{\binom{12}{3}}.$$

The distribution is not binomial because the draws are not independent.

Problem 8

Let I_j be the indicator that the jth person in a random permutation of $\{1, 2, \ldots, n\}$ is a fixed point. Use indicators to compute the expected number of fixed points.

Sketch. Let

$$F = \sum_{j=1}^{n} I_j.$$

Then

$$\mathbb{E}[F] = \sum_{j=1}^{n} \mathbb{E}[I_j] = \sum_{j=1}^{n} \mathbb{P}(I_j = 1).$$

For each j, the probability that position j contains j is $1/n$. Hence

$$\mathbb{E}[F] = n \cdot \frac{1}{n} = 1.$$

The indicator method avoids the harder task of finding the full distribution of F.

Problem 9

If X and Y are independent with $\mathbb{E}[X] = 2$, $\mathbb{E}[Y] = 3$, $\mathrm{Var}(X) = 4$, and $\mathrm{Var}(Y) = 5$, compute $\mathbb{E}[2X - 3Y]$ and $\mathrm{Var}(2X - 3Y)$.

Sketch. By linearity,

$$\mathbb{E}[2X - 3Y] = 2\mathbb{E}[X] - 3\mathbb{E}[Y] = 4 - 9 = -5.$$

By independence,

$$\mathrm{Var}(2X - 3Y) = 4\mathrm{Var}(X) + 9\mathrm{Var}(Y) = 16 + 45 = 61.$$

Without independence one would need the covariance term.

Problem 10

Show that if $X \geq 0$ and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Then apply this with $X = Y^2$ to derive Chebyshev's inequality.

Sketch. Markov's inequality follows because $X \geq a\mathbf{1}_{\{X \geq a\}}$, hence

$$\mathbb{E}[X] \geq a\mathbb{P}(X \geq a).$$

Now set $X = (Y - \mu)^2$ and $a = t^2$ to get

$$\mathbb{P}(|Y - \mu| \geq t) = \mathbb{P}((Y - \mu)^2 \geq t^2) \leq \frac{\mathrm{Var}(Y)}{t^2}.$$

This is Chebyshev's inequality.

C.3 Part II: Conditioning, transforms, and asymptotics

Problem 11

Suppose X and Y are discrete with joint pmf

$$p_{X,Y}(x, y) = c(x + y), \qquad x, y \in \{1, 2\}.$$

Find c, the marginal pmf of X, and $\mathbb{E}[Y \mid X = 1]$.

Sketch. Sum over all four possible pairs:

$$1 = c[(1 + 1) + (1 + 2) + (2 + 1) + (2 + 2)] = c(2 + 3 + 3 + 4) = 12c,$$

so $c = 1/12$. Then

$$\mathbb{P}(X = 1) = \frac{2 + 3}{12} = \frac{5}{12}, \qquad \mathbb{P}(X = 2) = \frac{3 + 4}{12} = \frac{7}{12}.$$

Also,

$$\mathbb{P}(Y = 1 \mid X = 1) = \frac{2/12}{5/12} = \frac{2}{5}, \qquad \mathbb{P}(Y = 2 \mid X = 1) = \frac{3/12}{5/12} = \frac{3}{5}.$$

Hence

$$\mathbb{E}[Y \mid X = 1] = 1 \cdot \frac{2}{5} + 2 \cdot \frac{3}{5} = \frac{8}{5}.$$

Problem 12

Let X and Y be independent exponential random variables with rate 1. Find the density of $X + Y$.

Sketch. Use convolution:

$$f_{X+Y}(t) = \int_0^t e^{-x} e^{-(t-x)}\, dx = \int_0^t e^{-t}\, dx = te^{-t}, \qquad t > 0.$$

This is the Gamma$(2, 1)$ density. The interpretation is that sums of independent exponentials of the same rate are Gamma.

Problem 13

Let $X \sim \text{Unif}(0, 1)$ and $Y = -\log X$. Find the distribution of Y.

Sketch. Use the cdf method. For $y \geq 0$,

$$\mathbb{P}(Y \leq y) = \mathbb{P}(-\log X \leq y) = \mathbb{P}(X \geq e^{-y}) = 1 - e^{-y}.$$

So $Y \sim \text{Exp}(1)$. This transformation appears often in simulation via inverse cdfs.

Problem 14

Suppose X is integrable and Y is discrete. Explain why

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y).$$

Sketch. This is the law of total expectation. Write

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X\mathbf{1}_{\{Y=y\}}]$$

and then use

$$\mathbb{E}[X\mathbf{1}_{\{Y=y\}}] = \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y).$$

Conceptually, conditioning partitions the sample space into simpler pieces and then averages over the pieces.

Problem 15

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Show that

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Sketch. Use independence:

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

This is the variance identity behind the weak law and the CLT.

Problem 16

State the weak law of large numbers and prove it using Chebyshev's inequality under the finite-variance assumption.

Sketch. The statement is

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu$$

for i.i.d. X_i with mean μ and finite variance. By the previous problem,

$$\operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Chebyshev then gives

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

Problem 17

Let $X_n \sim \operatorname{Bin}(n, \lambda/n)$. Use generating functions to show $X_n \xrightarrow{d} \operatorname{Pois}(\lambda)$.

Sketch. The pgf is

$$G_n(s) = \left(1 + \frac{\lambda}{n}(s-1)\right)^n.$$

As $n \to \infty$,

$$G_n(s) \to e^{\lambda(s-1)},$$

which is the pgf of the Poisson(λ) distribution. Hence X_n converges in distribution to Poisson(λ).

Problem 18

Suppose $X_i \sim \operatorname{Ber}(p)$ independently. Use the CLT to approximate

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \le p + \frac{a}{\sqrt{n}}\right)$$

for large n.

Sketch. Let $\hat{p}_n = \overline{X}_n$. Then

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} Z.$$

So for large n,

$$\mathbb{P}\left(\hat{p}_n \le p + \frac{a}{\sqrt{n}}\right) \approx \Phi\left(\frac{a}{\sqrt{p(1-p)}}\right).$$

The problem is crafted to make the CLT scaling visible.

Problem 19

Suppose $X_n \xrightarrow{a.s.} X$ and $|X_n| \le Y$ for all n, where $\mathbb{E}[Y] < \infty$. Show that $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Sketch. This is the dominated convergence theorem. Almost sure convergence gives pointwise convergence outside a null set, and domination by an integrable variable allows the limit to pass through expectation.

Problem 20

Let X be nonnegative integer-valued. Show that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \ge k).$$

Sketch. Write

$$X = \sum_{k=1}^{\infty} \mathbf{1}_{\{X \ge k\}}.$$

Taking expectations and using monotone convergence or term-by-term linearity gives the identity. This formula is especially useful for waiting-time distributions.

C.4 Part III: Poisson processes, Markov chains, and further topics

Problem 21

Customers arrive according to a Poisson process of rate 4 per hour. What is the probability of exactly three arrivals in the first half hour? What is the probability of no arrivals in the first half hour?

Sketch. In a half hour the mean count is $\lambda t = 4 \cdot 0.5 = 2$. Therefore

$$\mathbb{P}(N(0.5) = 3) = e^{-2}\frac{2^3}{3!}, \qquad \mathbb{P}(N(0.5) = 0) = e^{-2}.$$

Always translate the time interval first into the correct Poisson parameter.

Problem 22

For a rate-λ Poisson process, identify the distribution of the waiting time to the first arrival and compute its mean.

Sketch. If T is the first arrival time,

$$\mathbb{P}(T > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t},$$

so $T \sim \mathrm{Exp}(\lambda)$. Its mean is $1/\lambda$.

Problem 23

Let N_1 and N_2 be independent Poisson processes with rates 2 and 5. What is the rate of the superposed process $N_1 + N_2$? If each arrival in a rate-7 Poisson process is independently kept with probability $1/3$, what is the rate of the retained process?

Sketch. Superposition adds rates, so the first answer is 7. Thinning multiplies the rate by the keep probability, so the second answer is $7/3$.

Problem 24

Consider the Markov chain on states $\{0, 1\}$ with transition matrix

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}.$$

Find the stationary distribution.

Sketch. Let $\pi = (\pi_0, \pi_1)$. Solve

$$\pi_0 = 0.9\pi_0 + 0.4\pi_1, \qquad \pi_0 + \pi_1 = 1.$$

The first equation becomes $0.1\pi_0 = 0.4\pi_1$, so $\pi_0 = 4\pi_1$. Hence $5\pi_1 = 1$, giving

$$\pi_1 = \frac{1}{5}, \qquad \pi_0 = \frac{4}{5}.$$

Problem 25

A fair random walk on $\{0, 1, 2, 3, 4\}$ is absorbed at 0 and 4. Starting from 2, what is the probability of hitting 4 before 0?

Sketch. For fair gambler's ruin,

$$\mathbb{P}_i(\tau_4 < \tau_0) = \frac{i}{4}.$$

Thus starting from 2 the probability is $1/2$.

Problem 26

In the same chain, what is the expected time to absorption starting from 2?

Sketch. For fair gambler's ruin on $\{0, 1, \dots, N\}$,

$$\mathbb{E}_i[T] = i(N - i).$$

With $i = 2$ and $N = 4$, the expected time is

$$2(4 - 2) = 4.$$

A first-step recursion or the martingale $S_n^2 - n$ both lead to the same answer.

Problem 27

Suppose X_n is a Markov chain with transition matrix P, and f is a function on the state space satisfying $Pf = f$. Show that $f(X_n)$ is a martingale.

Sketch. Using the Markov property,

$$\mathbb{E}[f(X_{n+1}) \mid X_n = i] = \sum_j p_{ij} f(j) = (Pf)(i) = f(i).$$

Hence

$$\mathbb{E}[f(X_{n+1}) \mid X_n] = f(X_n).$$

Since conditioning on the full past reduces to conditioning on the present state, this gives the martingale property with respect to the natural filtration.

Problem 28

A branching process starts with one ancestor. Each individual has 0 or 2 children with probabilities 0.6 and 0.4. Compute the mean offspring number and determine whether extinction occurs almost surely.

Sketch. The offspring mean is

$$m = 0 \cdot 0.6 + 2 \cdot 0.4 = 0.8 < 1.$$

Since the process is subcritical, extinction occurs almost surely.

Problem 29

For the same branching process, compute the extinction probability directly from the generating function.

Sketch. The offspring pgf is

$$f(s) = 0.6 + 0.4s^2.$$

The extinction probability solves

$$q = 0.6 + 0.4q^2.$$

One solution is $q = 1$. Because the process is subcritical, the smallest solution in $[0, 1]$ is indeed $q = 1$.

Problem 30

Let $M_n = X_1 + \cdots + X_n$ where the X_i are i.i.d. with mean 0 and finite variance. Explain why M_n is a martingale and why M_n^2 is usually not a martingale, even though $M_n^2 - n\mathrm{Var}(X_1)$ is.

Sketch. The first claim follows from independence and mean zero:

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n + \mathbb{E}[X_{n+1}] = M_n.$$

For the square,

$$\mathbb{E}[M_{n+1}^2 \mid \mathcal{F}_n] = M_n^2 + \mathrm{Var}(X_1),$$

so there is a positive drift. Subtracting $n\mathrm{Var}(X_1)$ removes that drift, producing a martingale.

C.5 A few final meta-strategies

Before ending the review appendix, it is worth recording several habits that distinguish strong probability solutions from merely correct ones.

1. Define the random variable early

If a problem asks for an expectation, define the variable whose expectation you want. If it asks for a hitting probability, define that probability as a function of the starting state. Naming the object often reveals the method.

2. Write conditioning in full once

When using Bayes, total expectation, or first-step analysis, write the conditioning decomposition explicitly at least once. This reduces sign errors and missing terms.

3. State the approximation regime

If you use a Poisson or normal approximation, explain why. Is the count rare? Is the sample size large? Is a continuity correction appropriate? The approximation itself is only half the answer.

4. Check edge cases

A formula that fails in a boundary case usually signals an algebra or modeling mistake. For example, a probability should remain in $[0, 1]$, and an expected time should not become negative.

5. Look for multiple routes

Many good problems admit at least two solutions: counting versus indicators, recursion versus martingales, convolution versus transforms. Comparing routes is one of the best ways to deepen understanding.

C.6 Summary

These review problems are designed to be cumulative and method-oriented.

- Part I emphasizes counting, distributions, expectation, and inequalities.
- Part II emphasizes conditioning, transforms, convergence, and approximation.
- Part III emphasizes stochastic processes and the interplay between recursion and martingale thinking.

A student who can solve most of these problems cleanly is not only ready for exams in a course like Stat 134, but also ready to begin more advanced probability, statistics, or stochastic-process work with confidence.

References and Further Reading

[1] Jean Jacod and Philip Protter, *Probability Essentials*, 2nd ed., Springer, 2004.

[2] Rick Durrett, *Probability: Theory and Examples*, 4th ed., Cambridge University Press, 2010.

[3] James Blitzstein and Jessica Hwang, *Introduction to Probability*, 2nd ed., CRC Press, 2019.

[4] Geoffrey Grimmett and David Stirzaker, *Probability and Random Processes*, 4th ed., Oxford University Press, 2020.

[5] David Williams, *Probability with Martingales*, Cambridge University Press, 1991.

[6] Jim Pitman, *Probability*, Springer, 1993.