Part IA
Probability

Prof Perla Sousi
Lent 2026
Version 20260523

These are Zixuan's notes for **Part IA – Probability** at the University of Cambridge in 2026. The notes are not endorsed by the lecturers or the University, and all errors are my own.

The latest version of this document is available at academic.micfong.space. Please direct any comments to my CRSid email or use the contact details listed on the site.

This document is typeset using Typst. All figures are created using Inkscape.

Contents

Syllabus and Overview **4**

1 Basic Concepts **5**
- 1.1 Probability Spaces 5
- 1.2 Combinatorial Analysis 7
- 1.3 Stirling's Formula 8

2 Axiomatic Approach **12**
- 2.1 Probability Measure 12
- 2.2 Inclusion-Exclusion Formula 13
- 2.3 Independence 16
- 2.4 Conditional Probability 17
- 2.5 Discrete Probability Distributions 22

3 Discrete Random Variables **26**
- 3.1 Expectation 27
- 3.2 Variance and Covariance 32
- 3.3 Inequalities 37
- 3.4 Multiple Discrete Random Variables 40
  - 3.4.1 Joint Distribution and Conditional Distribution 40
  - 3.4.2 Distribution of the Sum of Random Variables 41
  - 3.4.3 Conditional Expectation 42
- 3.5 Random Walks 46
  - 3.5.1 Simple Random Walk 46
  - 3.5.2 Time to Absorption 47
- 3.6 Probability Generating Functions 48
  - 3.6.1 Introduction 48
  - 3.6.2 Sum of a Random Number of Random Variables 51
- 3.7 Branching Processes 52
  - 3.7.1 Generating Functions of Branching Processes 53
  - 3.7.2 Extinction Probability 54

4 Continuous Random Variables **58**
- 4.1 Probability Distribution Function 58
- 4.2 Uniform Distribution 60
- 4.3 Exponential Distribution 60
- 4.4 Expectation and Variance of a Continuous Random Variable 62
- 4.5 Normal Distribution 64
  - 4.5.1 Introduction 64
  - 4.5.2 Linear Transformations of Normal Distributions 66

4.6 Multivariate Density Functions 68
4.6.1 Introduction 68
4.6.2 Marginal Density Functions 69
4.6.3 Sum of Independent Random Variables 70
4.6.4 Conditional Density Functions 70
4.6.5 Transformation of Random Variables 71
4.7 Order Statistics of a Random Sample 72
4.7.1 Order Statistics of Exponential Distributions 73
4.8 Moment Generating Functions 73
4.8.1 Gamma Distribution 74
4.8.2 MGF of the Normal Distribution 75
4.8.3 Multivariate Moment Generating Functions 76
4.9 Multidimensional Gaussian Random Variables 77
4.9.1 Introduction 77
4.9.2 Construction of a Gaussian Random Vector 79
4.9.3 Density of a Gaussian Vector 80
4.9.4 Bivariate Gaussian Distribution 82

5 Convergence Results and Limit Theorems 84
5.1 Convergence Results 84
5.2 Central Limit Theorem 87
5.2.1 Sampling Error via the CLT 91
5.3 Simulation of Random Variables 92
5.3.1 Box-Muller Transform 92
5.4 Rejection Sampling 93

Syllabus and Overview

Lent Term, 2026 [24 Lectures]

Basic Concepts [3 Lectures]

Classical probability, equally likely outcomes. Combinatorial analysis, permutations and combinations. Stirling's formula (asymptotics for $\log n!$ proved).

Axiomatic Approach [5 Lectures]

Axioms (countable case). Probability spaces. Inclusion-exclusion formula. Continuity and subadditivity of probability measures. Independence. Binomial, Poisson and geometric distributions. Relation between Poisson and binomial distributions. Conditional probability, Bayes's formula. Examples, including Simpson's paradox.

Discrete Random Variables [7 Lectures]

Expectation. Functions of a random variable, indicator function, variance, standard deviation. Covariance, independence of random variables. Generating functions: sums of independent random variables, random sum formula, moments.

Conditional expectation. Random walks: gambler's ruin, recurrence relations. Difference equations and their solution. Mean time to absorption. Branching processes: generating functions and extinction probability. Combinatorial applications of generating functions.

Continuous Random Variables [6 Lectures]

Distributions and density functions. Expectations; expectation of a function of a random variable. Uniform, normal and exponential random variables. Memoryless property of exponential distribution.

Joint distributions: transformation of random variables (including Jacobians), examples. Simulation: generating continuous random variables, Box–Muller transform, rejection sampling. Geometrical probability: Bertrand's paradox, Buffon's needle. Correlation coefficient, bivariate normal random variables.

Inequalities and Limits [3 Lectures]

Markov's inequality, Chebyshev's inequality. Weak law of large numbers. Convexity: Jensen's inequality for general random variables, AM/GM inequality.

Moment generating functions and statement (no proof) of continuity theorem. Statement of central limit theorem and sketch of proof. Examples, including sampling.

Lecture 1 · 2026-01-23

1 Basic Concepts

We shall begin by general definitions in probability theory.

1.1 Probability Spaces

Definition 1.1 (Probability Space)

Suppose Ω is a set and $\mathcal{F}$ is a collection of subsets of Ω. Then we call $\mathcal{F}$ a **σ-algebra** if

- $\Omega \in \mathcal{F}$ [the whole space is in $\mathcal{F}$.]
- If $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$ [closed under complements.]
- If $(A_n)_{n\in\mathbb{N}}$ is a countable collection of sets in $\mathcal{F}$, then [closed under countable unions.]

$$\bigcup_n A_n \in \mathcal{F}.$$

Let $\mathcal{F}$ be a σ-algebra on Ω. A function

$$\mathbb{P} : \mathcal{F} \to [0,1]$$

is called a **probability measure** if

- $\mathbb{P}(\Omega) = 1$
- **Countable additivity.** For any countable disjoint collection of sets $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$, we have

$$\mathbb{P}\left(\bigcup_n A_n\right) = \sum_n \mathbb{P}(A_n).$$

We say $\mathbb{P}(A)$ to be the **probability** of the event $A \in \mathcal{F}$.

We call $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability space**.

Definition 1.2 (Outcomes and Events)

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

The elements of Ω are called **outcomes**, and the elements of $\mathcal{F}$ are called **events**.

Remark. We talk about probabilities of events instead of outcomes.

When Ω is countable, we take $\mathcal{F}$ to be the power set of Ω.

Proposition 1.3 (Properties of Probability Measures)

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Then for any $A, B \in \mathcal{F}$, we have

- $\mathbb{P}\left(A^{\complement}\right) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\varnothing) = 0$
- $\mathbb{P}(A) \leqslant \mathbb{P}(B)$ if $A \subseteq B$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Example 1.4

1. Consider rolling a fair die. We have
$$\begin{aligned} \Omega &= \{1,2,3,4,5,6\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \mathbb{P}(\{\omega\}) &= \frac{1}{6} && \forall \omega \in \Omega. \\ \mathbb{P}(A) &= \frac{|A|}{6} && \forall A \in \mathcal{F}. \end{aligned}$$

2. Consider
$$\begin{aligned} \Omega &= \{\omega_1, \ldots, \omega_n\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \mathbb{P}(A) &= \frac{|A|}{|\Omega|} && \forall A \in \mathcal{F}. \end{aligned}$$
This models the experiment of picking a uniformly random element from Ω. We have
$$\mathbb{P}\left(\{\omega\} = \frac{1}{|\Omega|}\right) \quad \forall \omega \in \Omega.$$

3. Consider picking balls from a bag. Suppose we have n balls labelled $\{1, 2, \ldots, n\}$. We pick k balls at random without replacement. Then we have
$$\begin{aligned} \Omega &= \{A \subseteq \{1, \ldots, n\} : |A| = k\} \\ |\Omega| &= \binom{n}{k} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \mathbb{P}(\{\omega\}) &= \frac{1}{\binom{n}{k}} && \forall \omega \in \Omega. \end{aligned}$$

4. Consider a well-shuffled deck of 52 cards. Then
$$\begin{aligned} \Omega &= \{\text{all permutations of 52 cards}\} \\ \mathbb{P}(\text{top 2 cards are aces}) &= \frac{4 \cdot 3 \cdot 50!}{52!} = \frac{1}{221}. \end{aligned}$$

5. **Largest digit problem.** Consider a string of n random digits from $0, \ldots, 9$. Then
$$\begin{aligned} \Omega &= \{0, 1, \ldots, 9\}^n \\ \mathcal{F} &= \mathcal{P}(\Omega). \end{aligned}$$

Then we define

$$A_k = \{\text{no digit exceeds } k\}.$$

And so

$$\mathbb{P}(A_k) = \frac{(k+1)^n}{10^n}.$$

Another example is the event

$$B_k = \{\text{largest digit is } k\}.$$

Then

$$\mathbb{P}(B_k) = \mathbb{P}(A_k \setminus A_{k-1}) = \frac{(k+1)^n - (k)^n}{10^n}.$$

6. **Birthday problem.** There are n people. What is the probability that at least two people share a birthday?

   We may assume that there are 365 days in a year, and each person's birthday is equally likely to be any of the 365 days, independently of other people.

   We have

   $$\Omega = \{1, 2, ..., 365\}^n$$
   $$\mathcal{F} = \mathcal{P}(\Omega).$$

   Then

   $$\mathbb{P}(\{\omega\}) = \frac{1}{365^n} \quad \forall \omega \in \Omega.$$

   Let A be the event that at least two people share a birthday. Then we have

   $$\mathbb{P}\left(A^{\complement}\right) = \frac{365}{365} \cdot \frac{364}{365} \cdot ... \cdot \frac{365 - n + 1}{365}.$$

   Thus

   $$\mathbb{P}(A) = 1 - \mathbb{P}\left(A^{\complement}\right).$$

1.2 Combinatorial Analysis

1. Let Ω be a finite set with $|\Omega| = n$.

   We want to partition Ω into k disjoint subsets $\Omega_1, ..., \Omega_k$ with $|\Omega_i| = n_i$ and $\sum_{n_i} = n$. Consider the number of ways to do this.

   Let M be the number of ways to do this. Then

   $$M = \binom{n}{n_1} \cdot \binom{n - n_1}{n_2} \cdot ... \cdot \binom{n - n_1 - ... - n_{k-1}}{n_k} = \frac{n!}{n_{1!} \cdot n_{2!} \cdot ... \cdot n_{k!}} = \binom{n}{n_1, n_2, ..., n_k}.$$

2. Let $f : \{1, ..., k\} \to \{1, ..., n\}$. We say that f is

- strictly increasing if $x < y \Rightarrow f(x) < f(y)$
- increasing if $x < y \Rightarrow f(x) \leqslant f(y)$

Each strictly increasing functions can be determined by its range. Hence the number of strictly increasing functions is $\binom{n}{k}$.

Lecture 2 · 2026-01-26

For increasing functions, define a bijection

$$g : \{f : \{1, ..., k\} \to \{1, ..., n\} \text{ increasing}\} \to \{f : \{1, ..., k\} \to \{1, ..., n\} \text{ strictly increasing}\}$$

by

$$g(f(i)) = f(i) + i - 1.$$

Hence the number of increasing functions is $\binom{n+k-1}{k}$.

1.3 Stirling's Formula

Notation. Let (a_n), (b_n) be 2 sequences. We write

$$(a_n) \sim (b_n) \quad \text{as} \quad n \to \infty$$

if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$.

Theorem 1.5 (Stirling's Formula)

As $n \to \infty$, we have

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

We shall first consider a weaker statement.

Proposition 1.6

$$\log(n!) \sim n \log(n) \quad \text{as} \quad n \to \infty.$$

Proof. Define

$$I_n = \log(n!) = \log 2 + ... + \log(n).$$

For $x \in \mathbb{R}$, we shall write $\lfloor x \rfloor$ as the integer part of x. Then

$$\log\lfloor x \rfloor \leqslant \log x \leqslant \log\lfloor x + 1 \rfloor.$$

Integrating from 1 to n, we have

$$\begin{aligned} \sum_{k=1}^{n-1} \log k \leqslant & \int_1^n \log x \, dx \leqslant \sum_{k=2}^{n} \log k \\ I_{n-1} \leqslant & \; n \log n - n + 1 \leqslant I_n \\ n \log n - n + 1 \leqslant & \; I_n \leqslant (n+1)\log(n+1) - (n+1) + 1. \end{aligned}$$

Dividing by $n \log n$, we have

$$\frac{l_n}{n \log n} \to 1 \quad \text{as} \quad n \to \infty.$$

Proof. [of Stirling's Formula 1.5, non-examinable proof] For all $f : \mathbb{R} \to \mathbb{R}$ that is twice differentiable, for all $a < b$ we have

$$\int_a^b f(x)\,\mathrm{d}x = \frac{f(a)+f(b)}{2}(b-a) - \frac{1}{2}\int_a^b (x-a)(b-x)f''(x)\,\mathrm{d}x.$$

This can be checked using integration by parts twice.

Take $f(x) = \log x, a = k, b = k+1, k \in \mathbb{N}$. Then we get

$$\int_k^{k+1} \log x\,\mathrm{d}x = \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_k^{k+1} \frac{(x-k)(k+1-x)}{x^2}\,\mathrm{d}x.$$

Therefore,

$$\int_1^n \log x\,\mathrm{d}x = \frac{\log(n-1)! + \log(n!)}{2} + \sum_{k=1}^{n-1} a_k$$

where

$$a_k = \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x.$$

We have

$$n \log n - n + 1 = \log(n!) - \frac{\log n}{2} + \sum_{k=1}^{n-1} a_k$$

$$\log(n!) = n \log n - n + 1 + \frac{\log n}{2} - \sum_{k=1}^{n-1} a_k$$

$$n! = n^n \cdot \mathrm{e}^{-n} \cdot \sqrt{n} \cdot \exp\left(1 - \sum_{k=1}^{n-1} a_k\right).$$

Now, for a_k,

$$a_k = \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x \leqslant \frac{1}{2k^2}\int_0^1 x(1-x)\,\mathrm{d}x = \frac{1}{12k^2}.$$

So the series $\sum_{k=1}^{\infty} a_k$ converges. Let

$$A = \exp\left(1 - \sum_{k=1}^{\infty} a_k\right).$$

Therefore,

$$n! = A \cdot n^n \cdot \mathrm{e}^{-n} \cdot \sqrt{n} \cdot \underbrace{\exp\left(\sum_{k=n}^{\infty} a_k\right)}_{\to 1 \text{ as } n \to \infty}.$$

We have shown that

$$\frac{n!}{n^n \cdot \mathrm{e}^{-n}\sqrt{n}} \to A \text{ as } n \to \infty.$$

It remains to show that $A = \sqrt{2\pi}$. We already know that

$$n! \sim A \cdot n^n \cdot \mathrm{e}^{-n} \cdot \sqrt{n}.$$

Consider

$$\begin{aligned} 2^{-2n}\binom{2n}{n} &= 2^{-2n}\frac{(2n)!}{n!n!} \\ &\sim \frac{2^{-2n} \cdot (2n)^{2n} \cdot \mathrm{e}^{-2n}\sqrt{2n} \cdot A}{\left(n^n \cdot \mathrm{e}^{-n}\sqrt{n} \cdot A\right)^2} \\ &= \frac{\sqrt{2}}{A\sqrt{n}}. \end{aligned}$$

We shall use a different method to show that

$$2^{-2n} \cdot \binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}} \quad \text{as} \quad n \to \infty.$$

This would imply that $A = \sqrt{2\pi}$.

Consider

$$I_n = \int_0^{\frac{\pi}{2}} (\cos\theta)^n \, \mathrm{d}\theta \quad \text{with} \quad I_0 = \frac{\pi}{2}, I_1 = 1.$$

Using integration by parts, we have the recurrence relation

$$I_n = \left(\frac{n-1}{n}\right) I_{n-2}.$$

So

$$\begin{aligned} I_{2n} &= \frac{2n-1}{2n} I_{2n-2} \\ &= \frac{(2n-1)(2n-3) \cdot \ldots \cdot 1}{(2n)(2n-2) \cdot \ldots \cdot 2}\frac{\pi}{2} \\ &= \frac{(2n)!}{2^{2n}(n!)^2}\frac{\pi}{2} \\ &= \frac{\pi}{2}2^{-2n}\binom{2n}{n}. \end{aligned}$$

Thus $I_{2n} = \frac{\pi}{2} \cdot 2^{-2n}\binom{2n}{n}$.

Similarly,

$$I_{2n+1} = \frac{2n \cdot \ldots \cdot 4 \cdot 2}{(2n+1) \cdot \ldots \cdot 3 \cdot 1} I_1 = \frac{1}{2n+1}\left(2^{-2n} \cdot \binom{2n}{n}\right)^{-1}.$$

Now, if we have that $\frac{I_{2n}}{I_{2n+1}} \to 1$ as $n \to \infty$, then we have

$$\frac{\frac{\pi}{2} 2^{-2n}\binom{2n}{n}}{\frac{1}{2n+1}\left(2^{-2n} \cdot \binom{2n}{n}\right)^{-1}} \to 1 \quad \Rightarrow \quad \left(2^{-2n}\binom{2n}{n}\right)^2 \sim \frac{1}{\pi n} \quad \text{as} \quad n \to \infty.$$

In order to show that $\frac{I_{2n}}{I_{2n+1}} \to 1$ as $n \to \infty$, we note that

$$\frac{I_n}{I_{n-2}} \to 1 \text{ as } n \to \infty.$$

Recall that

$$I_n = \int_0^{\frac{\pi}{2}} (\cos\theta)^n \, d\theta.$$

Since I_n is decreasing in n,

$$\frac{I_{2n}}{I_{2n+1}} \leqslant \frac{I_{2n-1}}{I_{2n+1}} \to 1,$$
$$\frac{I_{2n}}{I_{2n+1}} \geqslant \frac{I_{2n}}{I_{2n+2}} \to 1.$$

So

$$\frac{I_{2n}}{I_{2n+1}} \to 1 \text{ as } n \to \infty.$$

This completes the proof.

Lecture 3 · 2025-01-28

2 Axiomatic Approach

Recall our definition of probability spaces in Definition 1.1.

2.1 Probability Measure

There are several important properties of the probability measure $\mathbb{P}$.

Proposition 2.1 (Countable Subaddivity)

If (A_n) is a collection in $\mathcal{F}$, *i.e.* $(A_n \in \mathcal{F})\ \forall n$, then

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \leqslant \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

Proof. Define $B_1 = A_1$ and for $n \geqslant 2$,

$$B_n = A_n \setminus (A_1 \cup \ldots \cup A_{n-1}).$$

Then (B_n) is a disjoint collection in $\mathcal{F}$ and $\bigcup_n A_n = \bigcup_n B_n$.

So

$$\mathbb{P}\left(\bigcup A_n\right) = \mathbb{P}\left(\bigcup B_n\right) = \sum_n \mathbb{P}(B_n) \leqslant \sum_n \mathbb{P}(A_n)$$

because $B_n \subseteq A_n$ for all n.

Proposition 2.2 (Continuity of Probability Measures 1)

Let $(A_n)_{n\in\mathbb{N}}$ where $A_n \in \mathcal{F}$ and $A_n \subseteq A_{n+1}$ for all n. We call (A_n) an **increasing sequence** in $\mathcal{F}$.
Then $\mathbb{P}(A_n)$ is an increasing sequence in $\mathbb{R}$ and converges to $\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n)$.

Proof. Define $B_1 = A_1$, and for $n \geqslant 2$,

$$B_n = A_n \setminus A_{n-1} = A_n \setminus (A_1 \cup \ldots \cup A_{n-1}).$$

Then (B_n) is a disjoint collection, and

$$\bigcup_{k=1}^{n} B_k = A_n$$

$$\mathbb{P}(A_n) = \bigcup_{k=1}^{n} B_k = \sum_{k=1}^{n} \mathbb{P}(B_k) \to \mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k\right) \quad \text{as} \quad n \to \infty.$$

Now we also have

$$\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k\right) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right).$$

Hence

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right).$$

Proposition 2.3 (Continuity of Probability Measures 2)

Let $(A_n)_{n\in\mathbb{N}}$ where $A_n \in \mathcal{F}$ and $A_n \supseteq A_{n+1}$ for all n. We call (A_n) a **decreasing sequence** in $\mathcal{F}$. Then $\mathbb{P}(A_n)$ is a decreasing sequence in $\mathbb{R}$ and converges to $\mathbb{P}(\bigcap_{n\in\mathbb{N}} A_n)$.

Proof. Taking complements, we have that $\left(A_n^{\complement}\right)$ is an increasing sequence in $\mathcal{F}$. By Continuity of Probability Measures 1 2.2, the result follows.

2.2 Inclusion-Exclusion Formula

Proposition 2.4 (Inclusion-Exclusion Formula)

Let $A, B \in \mathcal{F}$. Then

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

In general, for n events $A_1, \ldots, A_n \in \mathcal{F}$, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1\leqslant i_1<\ldots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right)\right).$$

Proof. We shall prove this by induction on n. The base case $n = 2$ has already been shown. Assume that the formula holds for $n-1$ events. We shall show it holds for n events.

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i \cup A_n\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} \underbrace{(A_i \cap A_n)}_{B_i}\right).
\end{aligned}$$

By induction hypothesis,

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) = \sum_{k=1}^{n-1} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n-1} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) \right)$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right) = \sum_{k=1}^{n-1} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n-1} \mathbb{P}\left(\underbrace{A_{i_1} \cap \ldots \cap A_{i_k} \cap A_n}_{B_{i_1} \cap \ldots \cap B_{i_k}} \right) \right).$$

Plugging these into the previous equation gives the result.

Lemma 2.5 (Bonferroni Inequalities)

Let $A, B \in \mathcal{F}$. We have

$$\mathbb{P}(A \cup B) \leqslant \mathbb{P}(A) + \mathbb{P}(B)$$

Let $A_1, \ldots, A_n$ be events in $\mathcal{F}$. Then for any $r \leqslant n$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \begin{cases} \leqslant \sum_{k=1}^{r} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) \right) & \text{if } r \text{ is odd} \\ \geqslant \sum_{k=1}^{r} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) \right) & \text{if } r \text{ is even} \end{cases}$$

Proof. We shall prove this by induction. The base case $n = 2$ is clear.

Suppose the result holds for $n - 1$ events. We shall show it holds for n events.

Assume that r is odd. Then

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i \cup A_n\right) &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} \underbrace{A_i \cap A_n}_{B_i}\right). \end{aligned}$$

By the induction hypothesis, since r is odd,

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) \leqslant \sum_{k=1}^{r} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n-1} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) \right)$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right) \geqslant \sum_{k=1}^{r-1} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n-1} \mathbb{P}\left(\underbrace{A_{i_1} \cap \ldots \cap A_{i_k} \cap A_n}_{B_{i_1} \cap \ldots \cap B_{i_k}} \right) \right).$$

Substituting these into the previous equation gives the result. The case where r is even is similar.

Remark. If Ω is a finite set and $(\omega, \mathcal{F}, \mathbb{P})$ is a probability space with $\mathcal{F} = \mathcal{P}(\Omega)$, and define

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \quad \forall A \in \mathcal{F}.$$

Take $A_1, ..., A_n \in \mathcal{F}$, then by the inclusion-exclusion formula on $\mathbb{P}(\bigcup_{i=1}^n A_i)$, we can get

$$\left|\bigcup_{i=1}^n A_i\right| = \sum_{k=1}^n (-1)^{k+1}\left(\sum_{1\leqslant i_1<...<i_k\leqslant n} \left|A_{i_1} \cap ... \cap A_{i_k}\right|\right).$$

This is the **inclusion-exclusion principle** in combinatorics.

Lecture 4 · 2025-01-30

Example 2.6 (Counting Surjections)

Let

$$\Omega = \{f : \{1, ..., n\} \to \{1, ..., m\}\}$$
$$A = \{f \in \omega : f \text{ is a surjection}\}.$$

We want to find $|A|$.

Define

$$A_i = \{f \in \Omega : i \notin f(\{1, ..., n\})\} \quad \forall i = 1, ..., m.$$

Then

$$A = A_1^\complement \cap ... \cap A_m^\complement = \left(\bigcup_{i=1}^m A_i\right)^\complement$$
$$|A| = |\Omega| - \left|\bigcup_{i=1}^m A_i\right|.$$

Note that we have

$$|\Omega| = m^n$$
$$\left|A_1 \cup ... \cup A_m\right| = \sum_{k=1}^m (-1)^{k+1}\left(\sum_{1\leqslant i_1<...<i_k\leqslant m} \left|A_{i_1} \cap ... \cap A_{i_k}\right|\right)$$
$$= \sum_{k=1}^m (-1)^{k+1}\binom{m}{k}(m-k)^n.$$

So

$$|A| = m^n - \sum_{k=1}^m (-1)^{k+1}\binom{m}{k}(m-k)^n.$$

Example 2.7 (Counting Derangements)

A derangement is a permutation with no fixed points. Let

$$\Omega = \{\text{permutations of } \{1, 2, \ldots, n\}\}$$
$$A = \{\text{derangement}\} = \{f \in \Omega : f(i) \neq i \quad \forall i = 1, \ldots, n\}.$$

Pick a permutation at random. Consider the probability that it is in A. For $i = 1, \ldots, n$, let

$$A_i = \{f \in \Omega : f(i) = i\}.$$

Then

$$A = A_1^{\complement} \cap \ldots \cap A_n^{\complement} = \left(\bigcup_{i=1}^{n} A_i\right)^{\complement}.$$

So

$$\mathbb{P}(A) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right).$$

Using Inclusion-Exclusion Formula 2.4, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \underbrace{\mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right)}_{= \frac{(n-k)!}{n!}} \right) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!}.
\end{aligned}$$

Therefore,

$$\mathbb{P}(A) = 1 - \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!} = \sum_{k=0}^{n} (-1)^{k} \frac{1}{k!}.$$

Note that as $n \to \infty$, $\mathbb{P}(A) \to \mathrm{e}^{-1}$.

2.3 Independence

Definition 2.8 (Independence of Events)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B \in \mathcal{F}$. We say that A and B are **independent** if

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

We write $A \perp B$.

Let $(A_n)_{n \in \mathbb{N}}$ be a collection of events in $\mathcal{F}$. We say that (A_n) are **mutually independent** if for any finite subset $I \subseteq \mathbb{N}$,

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

Remark. Pairwise independence does not imply mutual independence.

Example 2.9

Consider tossing a fair coin twice. Let

$$\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$$
$$\mathbb{P}(\{\omega\}) = \frac{1}{4} \quad \forall \omega \in \Omega.$$

Define events

$$A = \{(0,0),(0,1)\}$$
$$B = \{(0,0),(1,0)\}$$
$$C = \{(1,0),(0,1)\}.$$

Then we have

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2},$$
$$\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{4},$$
$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\varnothing) = 0.$$

Thus, A, B, C are pairwise independent but not mutually independent.

Proposition 2.10

If $A \perp B$, then $A \perp B^{\complement}$.

Proof.

$$\begin{aligned}\mathbb{P}\left(A \cap B^{\complement}\right) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) \quad \text{since } A \perp B \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A)\mathbb{P}\left(B^{\complement}\right).\end{aligned}$$

2.4 Conditional Probability

Definition 2.11 (Conditional Probability)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B \in \mathcal{F}$ where $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is

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

In particular, if $A \perp B$, then

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

Proposition 2.12 (Countable Additivity for Conditional Probability)

Let $(A_n)_{n\in\mathbb{N}}$ is a disjoint sequence in $\mathcal{F}$. Then for some $B \in \mathcal{F}$ where $\mathbb{P}(B) > 0$,

$$\mathbb{P}\left(\bigcup_n A_n \;\middle|\; B\right) = \sum_n \mathbb{P}(A_n \mid B).$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\bigcup_n A_n \;\middle|\; B\right) &= \frac{\mathbb{P}((\bigcup_n A_n) \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(\bigcup_n (A_n \cap B))}{\mathbb{P}(B)} \\
&= \frac{\sum_n \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \quad \text{by countable additivity} \\
&= \sum_n \left(\frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)}\right) \\
&= \sum_n \mathbb{P}(A_n \mid B).
\end{aligned}$$

Proposition 2.13 (Law of Total Probability)

Suppose $(B_n)_{n\in\mathbb{N}}$ is a disjoint collection if $\mathcal{F}$ such that $\bigcup_n B_n = \Omega$ and $\mathbb{P}(B_n) > 0$ for all n. Then for any $A \in \mathcal{F}$,

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

Proof.

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(A \cap \left(\bigcup_n B_n\right)\right) \\ &= \mathbb{P}\left(\bigcup_n (A \cap B_n)\right) \\ &= \sum_n \mathbb{P}(A \cap B_n) \quad \text{since } B_n \text{ are disjoint} \\ &= \sum_n \mathbb{P}(A \mid B_n)\mathbb{P}(B_n).\end{aligned}$$

Proposition 2.14 (Bayes' Formula)

Consider (B_n) to be a collection of disjoint events in $\mathcal{F}$ such that $\bigcup_n B_n = \Omega$ and $\mathbb{P}(B_n) > 0$ for all n. Then for any $A \in \mathcal{F}$ where $\mathbb{P}(A) > 0$,

$$\mathbb{P}(B_k \mid A) = \frac{\mathbb{P}(A \mid B_k)\mathbb{P}(B_k)}{\sum_n \mathbb{P}(A \mid B_n)\mathbb{P}(B_n)}.$$

Proof. By Conditional Probability 2.11,

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

By Law of Total Probability 2.13, we have

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

Also,

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

Plugging these into the first equation gives the result.

This formula is the basis of Bayesian statistics; if we know the probabilities of $\mathbb{P}(B_k)$ and we have a model which gives us $\mathbb{P}(A \mid B_k)$, then we can compute the posterior probability $\mathbb{P}(B_n \mid A)$.

Example 2.15 (False Positives for a Rare Condition)

Suppose that a rare condition A affects 0.1% of the population, *i.e.* $\mathbb{P}(A) = 0.001$. A test for the condition has a 98% true positive rate and a 1% false positive rate. [For affected individuals, the test is positive with probability 0.98, and for unaffected individuals, the test is positive with probability 0.01.]

An individual takes the test and tests positive. Consider the probability that they actually have the condition.

Define

$$A = \{\text{individual suffers from } A\}$$
$$P = \{\text{test is positive}\}.$$

We have

$$\mathbb{P}(A) = 0.001, \quad \mathbb{P}(P \mid A) = 0.98, \quad \mathbb{P}\left(P \;\middle|\; A^{\complement}\right) = 0.01.$$

Then

$$\mathbb{P}(A \mid P) = \frac{\mathbb{P}(P \mid A)\mathbb{P}(A)}{\mathbb{P}(P \mid A)\mathbb{P}(A) + \mathbb{P}\left(P \;\middle|\; A^{\complement}\right)\mathbb{P}\left(A^{\complement}\right)} = \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.01 \times 0.999} \approx 0.09.$$

Note that we can rewrite this as

$$\mathbb{P}(A \mid P) = \frac{1}{1 + \frac{\mathbb{P}(P|A^{\complement})\mathbb{P}(A^{\complement})}{\mathbb{P}(P|A)\mathbb{P}(A)}} = \frac{1}{1 + \frac{0.01 \times 0.999}{0.98 \times 0.001}} \approx 0.09.$$

Since $\mathbb{P}\left(A^{\complement}\right)$ and $\mathbb{P}(P \mid A)$ are both very close to 1, so the relavent ration is approximately

$$\frac{\mathbb{P}\left(P \;\middle|\; A^{\complement}\right)}{\mathbb{P}(A)}.$$

But $\mathbb{P}\left(P \;\middle|\; A^{\complement}\right) \gg \mathbb{P}(A)$ and hence the posterior probability is still quite small.

Lecture 5 · 2026-02-02

Example 2.16

Consider the probability that a person have 2 boys given that,

1. They have exactly 2 children, one of whom is a boy.

   Since no further information is given, we assume that all outcomes are equally likely.

   Let

   $$\text{BG} = \{\text{elder child is a boy, younger child is a girl}\}$$
   $$\text{GB} = \{\text{elder child is a girl, younger child is a boy}\}$$
   $$\text{BB} = \{\text{both children are boys}\}.$$
   $$\text{GG} = \{\text{both children are girls}\}.$$

   We want to find

   $$\begin{aligned}\mathbb{P}(\text{BB} \mid \text{BB} \cup \text{BG} \cup \text{GB}) &= \frac{\mathbb{P}(\text{BB})}{\mathbb{P}(\text{BB} \cup \text{BG} \cup \text{GB})} \\ &= \frac{1}{4} / \frac{3}{4} \\ &= \frac{1}{3}.\end{aligned}$$

2. They have exactly 2 children, the elder of whom is a boy.

We want to find

$$\begin{aligned}\mathbb{P}(\text{BB} \mid \text{BB} \cup \text{BG}) &= \frac{\mathbb{P}(\text{BB})}{\mathbb{P}(\text{BB} \cup \text{BG})} \\ &= \frac{1}{4} / \frac{2}{4} \\ &= \frac{1}{2}.\end{aligned}$$

3. They have exactly 2 children, exactly one of them is a boy, who was born on a Thursday.

   Let T denote the event that a child is a boy born on a Thursday, and N denote the event that a child is a boy not born on a Thursday. We want to find

$$\begin{aligned}\mathbb{P}(\text{TT} \cup \text{TN} \cup \text{NT} \mid \text{TT} \cup \text{TN} \cup \text{NT} \cup \text{GT} \cup \text{TG}) &= \frac{\mathbb{P}(\text{TT} \cup \text{TN} \cup \text{NT})}{\mathbb{P}(\text{TT} \cup \text{TN} \cup \text{NT} \cup \text{GT} \cup \text{TG})} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} \cdot \frac{1}{7} + \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} \cdot \frac{6}{7} + \frac{1}{2} \cdot \frac{6}{7} \cdot \frac{1}{2} \cdot \frac{1}{7}}{\frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} \cdot \frac{1}{7} + \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} \cdot \frac{6}{7} \cdot 2} \\ &= \frac{13}{27}.\end{aligned}$$

Example 2.17 (Simpson's Paradox)

Suppose the following data is collected for applicants to a Cambridge college:

		Admitted	Rejected	Admission Rate
All Applicants	**State**	25	25	50%
	Independent	28	22	56%
London Schools	**State**	15	22	41%
	Independent	5	8	38%
Cambridge Schools	**State**	10	3	77%
	Independent	23	14	62%

Note that within each school type, state school applicants have a higher admission rate than independent school applicants. However, when considering all applicants, independent school applicants have a higher admission rate than state school applicants.

This phenomenon is called **confounding** in statistics. It arises when we aggregate data from disparate populations.

In terms of conditional probability, let

$$
\begin{aligned}
A &= \{\text{individual is admitted}\} \\
B &= \{\text{individual is from London}\} \\
B^{\complement} &= \{\text{individual is from Cambridge}\} \\
C &= \{\text{individual is from a state school}\} \\
C^{\complement} &= \{\text{individual is from an independent school}\}.
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathbb{P}(A \mid B \cap C) &> \mathbb{P}\left(A \;\middle|\; B \cap C^{\complement}\right), \\
\mathbb{P}\left(A \;\middle|\; B^{\complement} \cap C\right) &> \mathbb{P}\left(A \;\middle|\; B^{\complement} \cap C^{\complement}\right),
\end{aligned}
$$

but

$$
\mathbb{P}(A \mid C) < \mathbb{P}\left(A \;\middle|\; C^{\complement}\right).
$$

Note that

$$
\begin{aligned}
\mathbb{P}(A \mid C) &= \mathbb{P}(A \cap B \mid C) + \mathbb{P}\left(A \cap B^{\complement} \;\middle|\; C\right) \\
&= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}\left(A \cap B^{\complement} \cap C\right)}{\mathbb{P}(C)} \\
&= \frac{\mathbb{P}(A \mid B \cap C)\mathbb{P}(B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}\left(A \;\middle|\; B^{\complement} \cap C\right)\mathbb{P}\left(B^{\complement} \cap C\right)}{\mathbb{P}(C)} \\
&= \mathbb{P}(A \mid B \cap C)\mathbb{P}(B \mid C) + \mathbb{P}\left(A \;\middle|\; B^{\complement} \cap C\right)\mathbb{P}\left(B^{\complement} \;\middle|\; C\right) \\
&> \mathbb{P}\left(A \;\middle|\; B \cap C^{\complement}\right)\mathbb{P}(B \mid C) + \mathbb{P}\left(A \;\middle|\; B^{\complement} \cap C^{\complement}\right)\mathbb{P}\left(B^{\complement} \;\middle|\; C\right)
\end{aligned}
$$

Now, if we (falsely) assume that $\mathbb{P}(B \mid C) = \mathbb{P}\left(B \;\middle|\; C^{\complement}\right)$, then

$$
\begin{aligned}
\mathbb{P}(A \mid C) &> \mathbb{P}\left(A \;\middle|\; B \cap C^{\complement}\right)\mathbb{P}\left(B \;\middle|\; C^{\complement}\right) + \mathbb{P}\left(A \;\middle|\; B^{\complement} \cap C^{\complement}\right)\mathbb{P}\left(B^{\complement} \;\middle|\; C^{\complement}\right) \\
&= \mathbb{P}\left(A \;\middle|\; C^{\complement}\right).
\end{aligned}
$$

Nonetheless, this does not hold in our example.

2.5 Discrete Probability Distributions

Definition 2.18 (Discrete Probability Distribution)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is countable with

$$
\begin{aligned}
\Omega &= \{\omega_1, \omega_2, \ldots\} \\
\mathcal{F} &= \mathcal{P}(\Omega).
\end{aligned}
$$

If we know $\mathbb{P}(\{\omega_i\})$ for all i, then $\forall A \subseteq \Omega$,

$$\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\{\omega_i\}).$$

In this case, we say that $(\mathbb{P}(\{\omega_i\}))_i$ a **discrete probability distribution**.

We write $p_i = \mathbb{P}(\{\omega_i\})$ for all i.

Proposition 2.19 (Properties of Discrete Probability Distributions)

Let $(p_i)_i$ be a discrete probability distribution on a countable set $\Omega = \{\omega_1, \omega_2, ...\}$. Then

1. $p_i \geqslant 0$ for all i.
2. $\sum_i p_i = 1$.

Let us see some examples of discrete probability distributions.

Definition 2.20 (Bernoulli Distribution)

A **Bernoulli distribution** models the outcome of a single binary experiment, such as a biased coin toss.

The Bernoulli distribution $\mathrm{Ber}(p)$ with parameter $p \in [0, 1]$ is defined by

$$\begin{aligned} \Omega &= \{0, 1\} \\ p_0 &= 1 - p \\ p_1 &= p. \end{aligned}$$

Definition 2.21 (Binomial Distribution)

A **binomial distribution** models the number of successes in n independent Bernoulli trials, each with success probability $p \in [0, 1]$.

The binomial distribution $\mathrm{Bin}(n, p)$ with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ is defined by

$$\begin{aligned} \Omega &= \{0, 1, ..., n\} \\ p_k &= \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k = 0, ..., n. \end{aligned}$$

where (p_k) is called the binomial distribution. Note that in the case of a coin toss,

$$p_k = \mathbb{P}(\text{obtaining } k \text{ heads in } n \text{ tosses}).$$

Definition 2.22 (Multinomial Distribution)

A **multinomial distribution** models the number of occurrences of each outcome in n independent trials, each with k possible outcomes with probabilities $p_1, ..., p_k$.

The multinomial distribution $\mathrm{M}(n, p_1, ..., p_k)$ with parameters $n \in \mathbb{N}$ and $p_1, ..., p_k \in [0, 1]$ where $p_1 + ... + p_k = 1$ is defined by

$$\Omega = \left\{(n_1, \ldots, n_k) \in \mathbb{N}^k : \sum_{i=1}^{k} n_i = n\right\}$$
$$p_{n_1,\ldots,n_k} = \binom{n}{n_1, \ldots, n_k} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k} \quad \forall (n_1, \ldots, n_k) \in \Omega.$$

For example, suppose that there are k boxes and we throw n balls into these boxes independently, where each ball lands in box i with probability p_i. Then

$$\begin{aligned}\mathbb{P}(n_1 \text{ balls in box } 1, \ldots, n_k \text{ balls in box } k) &= \binom{n}{n_1} p_1^{n_1} \cdot \ldots \cdot \binom{n - n_1 - \ldots - n_{k-1}}{n_k} p_k^{n_k} \\ &= \binom{n}{n_1, \ldots, n_k} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k}.\end{aligned}$$

Lecture 6 · 2026-02-04

Definition 2.23 (Geometric Distribution)

A **geometric distribution** models the number of Bernoulli trials needed to get the first success, where each trial has success probability $p \in [0, 1]$.

The geometric distribution $\mathrm{Geo}(p)$ with parameter $p \in [0, 1]$ is defined by

$$\Omega = \mathbb{N}$$
$$p_k = (1 - p)^{k-1} p \quad \forall k \in \mathbb{N}.$$

Note that,

$$\sum_k p_k = \sum_k (1 - p)^{k-1} p = p \,/\, \frac{1}{1 - (1 - p)} = 1.$$

This models, for example, the number of times we need to toss a biased coin with heads probability p until we get the first heads.

Definition 2.24 (Poisson Distribution)

A **Poisson distribution** models the number of events occurring in a fixed interval of time or space, where these events occur with a known constant mean rate and independently of the time since the last event.

The Poisson distribution $\mathrm{Poi}(\lambda)$ with parameter $\lambda > 0$ is defined by

$$\Omega = \mathbb{Z}_{\geqslant 0}$$
$$p_k = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \quad \forall k \in \mathbb{N}_0.$$

Note that

$$\sum_k p_k = \sum_k \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} = \frac{\mathrm{e}^{-\lambda}}{\mathrm{e}^{\lambda}} = 1.$$

Proof. Consider the number of customers arriving at a shop in a time interval $[0, 1]$. We can discretise the time interval to $\left[\frac{i-1}{N}, \frac{i}{N}\right]$ with $i = 1, \ldots, N$, where each interval has a small probability p of a customer arriving. Then

$$\begin{aligned}
\mathbb{P}(k \text{ customers have arrived}) &= \binom{N}{k} p^k (1-p)^{N-k} \\
&= \frac{N!}{k!(N-k)!} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} \quad \text{with } p = \frac{\lambda}{N} \\
&= \frac{\lambda^k}{k!} \frac{N(N-1)\ldots(N-k+1)}{N^k} \left(1 - \frac{\lambda}{N}\right)^{N-k} \\
&\to \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \quad \text{as } N \to \infty.
\end{aligned}$$

3 Discrete Random Variables

Definition 3.1 (Random Variable)

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A **random variable** is a function $X : \Omega \to \mathbb{R}$, satisfying that $\forall x \in \mathbb{R}$,

$$\{\omega \in \Omega : X(\omega) \leqslant x\} \in \mathcal{F}.$$

We write $\{X \leqslant x\} = \{\omega \in \Omega : X(\omega) \leqslant x\}$ and $\forall A \subseteq \mathbb{R}$, $\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$.

Definition 3.2 (Indicator)

Let $A \in \mathcal{F}$. The **indicator** of A is a random variable $\mathbb{1}_A : \Omega \to [0, 1]$ defined by

$$\mathbb{1}_A(\omega) = \mathbb{1}(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^{\complement} \end{cases}$$

Definition 3.3 (Probability Distribution Function)

Suppose X is a random variable. Define the **probability distribution function** of X to be

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

where $F_X : \mathbb{R} \to [0, 1]$.

Definition 3.4 (Multidimensional Random Variable)

$X = (X_1, ..., X_n)$ is called a random variable in $\mathbb{R}^n$ if

$$(X_1, ..., X_n) : \Omega \to \mathbb{R}^n$$

is a function such that $\forall x_1, x_2, ..., x_n \in \mathbb{R}$,

$$\{X_1 \leqslant x_1, ..., X_n \leqslant x_n\} = \{\omega \in \Omega : X_1(\omega) \leqslant x_1, X_2(\omega) \leqslant x_2, ..., X_n(\omega) \leqslant x_n\} \in \mathcal{F}.$$

Equivalently, all X_i are real random variables.

Definition 3.5 (Discrete Random Variable)

A random variable X is called **discrete** if its range is countable.

Suppose it takes values in a countable set S, then for every $x \in S$ we write

$$p_x = \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\}).$$

We call $(p_x)_{x \in S}$ the **probability mass function** of X, or the **distribution** of X. Note that $\forall A \subseteq S$,

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_x.$$

Recall that two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definition 3.6 (Independence)

Let $X_1, X_2, ..., X_n$ be discrete random variables with values in $S_1, ..., S_n$. They are **independent** if for any $x_1 \in S_1, ..., x_n \in S_n$,

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdot ... \cdot \mathbb{P}(X_n = x_n).$$

Example 3.7

Consider tossing a p-coin N times independently. let $\Omega = \{0, 1\}$ with $\omega \in \Omega$, where

$$\omega = (\omega_1, \omega_2, ..., \omega_N)$$

and ω_i is the result of the i-th toss. Then

$$p_\omega = \mathbb{P}(\{\omega\}) = \prod_{k=1}^{N} p^{\omega_k}(1-p)^{1-\omega_k}.$$

For all $k = 1, ..., N$ define random variables $X_k : \Omega \to \{0, 1\}$ with $X_k(\omega) = \omega_k$. Then

$$\mathbb{P}(X_k = 1) = p$$
$$\mathbb{P}(X_k = 0) = 1 - p.$$

Note that X is a Bernoulli random variable with parameter p.

Claim. $X_1, ..., X_n$ are independent random variables.

Proof.

$$\begin{aligned}\mathbb{P}(X_1 = x_1, ..., X_n = x_n) &= \prod_{k=1}^{N} p^{x_k}(1-p)^{1-x_k} \\ &= \prod_{k=1}^{N} \mathbb{P}(X_k = x_k).\end{aligned}$$

Lecture 7 · 2026-02-06

3.1 Expectation

Definition 3.8 (Expectation for Non-Negative Discrete Random Variables)

Let Ω be a countable set and $X : \Omega \to \mathbb{R}$ be a discrete random variable.

For $X \geqslant 0$, the **expectation** of X is defined by

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\})X(\omega).$$

Alternatively, consider

$$\Omega_X = \{X(\omega) : \omega \in \Omega\}$$
$$\Omega = \bigcup_{x \in \Omega_X} \{\omega : X(\omega) = x\}$$

Then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} \sum_{\omega \in \{X=x\}} X(\omega) \cdot \mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} \sum_{\omega \in \{X=x\}} x \cdot \mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} x \sum_{\omega \in \{X=x\}} \mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} x\mathbb{P}(X = x).
\end{aligned}$$

So the expectation of X is the weighted average of the values taken by X, with weights given by the probabilities of X taking those values.

Example 3.9

Consder $X \sim \text{Bin}(N, p)$. Then $\forall k \in \{0, ..., N\}$,

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

Then, for the expectation,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{N} k \cdot \mathbb{P}(X = k) \\
&= \sum_{k=0}^{N} k \cdot \binom{N}{k} p^k (1-p)^{N-k} \\
&= \sum_{k=0}^{N} k \cdot \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} \\
&= \sum_{k=0}^{N-1} N \binom{N-1}{k} p^k (1-p)^{N-1-k} p \\
&= Np(p + 1 - p)^{N-1} \\
&= Np.
\end{aligned}$$

Example 3.10

Consider $X \sim \text{Poi}(\lambda)$ with $\lambda > 0$. Then for all $k \in \mathbb{Z}_{\geqslant 0}$,

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

Hence

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

Notation. Let X be a random variable. Define

$$X^+ = \max(X, 0) \quad \text{and} \quad X^- = \max(-X, 0).$$

Then

$$X = X^+ - X^- \quad \text{and} \quad |X| = X^+ + X^-.$$

Definition 3.11 (Expectation for General Discrete Random Variables)

Suppose X is discrete. We can define $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ as in Definition 3.8.

If at least one of $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ is finite, then we can define the **expectation** of X by

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

Otherwise, $\mathbb{E}[X]$ is not defined.

Definition 3.12 (Integrable Random Variable)

A random variable X is **integrable** if $\mathbb{E}[|X|] < \infty$.

Proposition 3.13

If $\mathbb{E}[X]$ is well-defined, then we have

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x\mathbb{P}(X = x).$$

Remark. We shall assume that whenever we write $\mathbb{E}[X]$, it is well-defined.

Proposition 3.14 (Properties of Expectation)

1. If $X \geqslant 0$, then $\mathbb{E}[X] \geqslant 0$.
2. If $X \geqslant 0$ and $\mathbb{E}[X] = 0$, then $\mathbb{P}(X = 0) = 1$.
3. If $c \in \mathbb{R}$, then $\mathbb{E}[cX] = c\mathbb{E}[X]$ and $\mathbb{E}[c + X] = c + \mathbb{E}[X]$.
4. If X and Y are random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
5. Let $c_1, \ldots, c_n \in \mathbb{R}$ and $X_1, \ldots, X_n$ be integrable random variables. Then
$$\mathbb{E}\left[\sum_{k=1}^{n} c_k X_k\right] = \sum_{k=1}^{n} c_k \mathbb{E}[X_k].$$
6. Suppose that $X_1, X_2, \ldots$ are non-negative random variables. Then
$$\mathbb{E}\left[\sum_{k=1}^{\infty} X_k\right] = \sum_{k=1}^{\infty} \mathbb{E}[X_k].$$

Proof. Suppose that Ω is countable.

For (6), we have

$$\begin{aligned}
\mathbb{E}\left[\sum_n X_n\right] &= \sum_{\omega\in\Omega}\left(\sum_n X_n(\omega)\right)\cdot \mathbb{P}(\{\omega\}) \\
&= \sum_n \sum_{\omega\in\Omega} X_n(\omega)\mathbb{P}(\{\omega\}) \quad \text{since all terms are non-negative} \\
&= \sum_n \mathbb{E}[X_n].
\end{aligned}$$

Example 3.15

Let $A \in \mathcal{F}, X = \mathbb{1}(A), X(\omega) = \mathbb{1}(\omega \in A)$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega\in\Omega} \mathbb{1}(\omega \in A)\cdot \mathbb{P}(\{\omega\}) \\
&= \sum_{\omega\in A} \mathbb{P}(\{\omega\}) \\
&= \mathbb{P}(A).
\end{aligned}$$

Proposition 3.16

Let $g : \mathbb{R} \to \mathbb{R}, X : \Omega \to \mathbb{R}$ and consider $g(X)$ defined by $g(X)(\omega) = g(X(\omega))$. Then $g(X)$ is a random variable, with

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

Proof. Let $Y = g(X)$. Then we have

$$\{Y = y\} = \{\omega : g(X(\omega)) = y\} = \left\{\omega : X(\omega) \in g^{-1}(\{y\})\right\} = \left\{X \in g^{-1}(\{y\})\right\}$$

Hence,

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y \in \Omega_y} y\mathbb{P}(Y = y) \\
&= \sum_{y \in \Omega_y} y\mathbb{P}\left(X \in g^{-1}(\{y\})\right) \\
&= \sum_{y \in \Omega_y} y \sum_{x \in g^{-1}(\{y\})} \mathbb{P}(X = x) \\
&= \sum_{y \in \Omega_Y} \sum_{x \in g^{-1}(\{y\})} g(x)\mathbb{P}(X = x) \\
&= \sum_{x \in \Omega_X} g(x)\mathbb{P}(X = x).
\end{aligned}$$

Proposition 3.17

Suppose $X \geqslant 0$ and takes integer values. Then

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geqslant k) = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

Proof. Note that for any $x \in \mathbb{N}$,

$$x = \sum_{k=1}^{\infty} \mathbb{1}(x \geqslant k) = \sum_{k=0}^{\infty} \mathbb{1}(x > k).$$

Since $X \geqslant 0$ and $X \in \mathbb{Z}_{\geqslant 0}$, we have

$$X(\omega) = \sum_{k=1}^{\infty} \mathbb{1}(X(\omega) \geqslant k) = \sum_{k=0}^{\infty} \mathbb{1}(X(\omega) > k).$$

Taking expectation on both sides, we get

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}(X \geqslant k)\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbb{1}(X > k)\right]$$

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{E}[\mathbb{1}(X \geqslant k)] = \sum_{k=0}^{\infty} \mathbb{E}[\mathbb{1}(X > k)]$$

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geqslant k) = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

Lecture 8 · 2026-02-09

With expectation, we can form another proof of the inclusion-exclusion formula.

Proof. [of Inclusion-Exclusion Formula 2.4] Let $A, B \in \mathcal{F}$. Then

$$\mathbb{1}\left(A^{\complement}\right) = 1 - 1(A)$$
$$\mathbb{1}(A \cap B) = \mathbb{1}(A) \cdot \mathbb{1}(B)$$
$$\mathbb{1}(A \cup B) = 1 - \mathbb{1}\left(A^{\complement} \cap B^{\complement}\right) = 1 - (1 - \mathbb{1}(A)) \cdot (1 - \mathbb{1}(B)).$$

More generally, if we have $A_1, ..., A_n \in \mathcal{F}$, then

$$\begin{aligned}\mathbb{1}(A_1 \cup ... \cup A_n) &= 1 - \mathbb{1}\left(A_1^{\complement} \cap ... \cap A_n^{\complement}\right) \\ &= 1 - \prod_{i=1}^{n}(1 - \mathbb{1}(A_i)) \\ &= \sum_{i-1}^{n} \mathbb{1}(A_i) - \sum_{1 \leqslant i < j \leqslant n} \mathbb{1}(A_i) \cdot \mathbb{1}\left(A_j\right) + ... + (-1)^{n-1}\mathbb{1}(A_1) \cdot ... \cdot \mathbb{1}(A_n).\end{aligned}$$

Taking expectation on both sides, since $\mathbb{E}[\mathbb{1}(A)] = \mathbb{P}(A)$, we get

$$\mathbb{P}(A_1 \cup ... \cup A_n) = \sum_{i-1}^{n} \mathbb{P}(A_i) - \sum_{1 \leqslant i < j \leqslant n} \mathbb{P}\left(A_i \cap A_j\right) + ... + (-1)^{n-1}\mathbb{P}(A_1 \cap ... \cap A_n).$$

3.2 Variance and Covariance

Definition 3.18 (Moment)

Let X be a random variable and $r \in \mathbb{N}$. We call $\mathbb{E}[X^r]$ the ***r*-th moment** of X, as long as it is well-defined.

Definition 3.19 (Variance)

The **variance** of X is defined by

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

The variance is a measure of concentration of the distribution of X around its expectation. The smaller the variance, the more concentrated the distribution is around its expectation.

Definition 3.20 (Standard Deviation)

The **standard deviation** of X is defined by $\sqrt{\operatorname{Var}(X)}$.

Proposition 3.21 (Propositions Of Var(X))

1. $\operatorname{Var}(X) \geqslant 0$.
2. If $\operatorname{Var}(X) = 0$, then $\mathbb{P}(X = \mathbb{E}[X]) = 1$.
3. If $c \in \mathbb{R}$, then $\operatorname{Var}(cX) = c^2 \operatorname{Var}(X)$ and $\operatorname{Var}(c + X) = \operatorname{Var}(X)$.

4. $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$.
5. $\mathrm{Var}(X) = \min_{c\in\mathbb{R}} \mathbb{E}\left[(X-c)^2\right]$, with the minimum attained at $c = \mathbb{E}[X]$.

Proof. For (4), we have

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

For (5), define $f(c) = \mathbb{E}\left[(X-c)^2\right]$. Then

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

Hence, taking derivative with respect to c, we get

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

Therefore, $f'(c) = 0$ if and only if $c = \mathbb{E}[X]$. Moreover, $f''(c) = 2 > 0$, so f is convex and the minimum is attained at $c = \mathbb{E}[X]$. Hence

$$\mathrm{Var}(X) = f(\mathbb{E}[X]) = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = \min_{c\in\mathbb{R}} \mathbb{E}\left[(X-c)^2\right].$$

Example 3.22

1. Consider $X \sim \mathrm{Bin}(n,p)$. Then $\mathbb{E}[X] = np$ and

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

Note that

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=2}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= n(n-1)p^2 \underbrace{\sum_{k=0}^{n-2} \frac{(n-2)!}{k!(n-2-k)!}p^k(1-p)^{n-2-k}}_{1} \\ &= n(n-1)p^2.\end{aligned}$$

Hence,

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

2. Consider $X \sim \mathrm{Poi}(\lambda)$. Then $\mathbb{E}[X] = \lambda$ and

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=2}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda^2 \underbrace{\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}}_{1} \\ &= \lambda^2.\end{aligned}$$

Hence,

$$\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Definition 3.23 (Covariance)

Let X and Y be random variables. The **covariance** of X and Y is defined by

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

It is a measure of the dependency between X and Y.

Proposition 3.24 (Properties of Covariance)

1. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
2. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
3. $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
4. Let $c \in \mathbb{R}$. Then $\mathrm{Cov}(cX,Y) = c\,\mathrm{Cov}(X,Y)$ and $\mathrm{Cov}(c+X,Y) = \mathrm{Cov}(X,Y)$.
5. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$.
6. Let $c \in \mathbb{R}$. Then $\mathrm{Cov}(c,X) = 0$.
7. Let X, Y, Z be random variables. Then $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.

   More generally, for all $c_1,\dots,c_n,d_1,\dots,d_n \in \mathbb{R}$, we have

$$\mathrm{Cov}\left(\sum_{i=1}^{n} c_k X_k, \sum_{j=1}^{n} d_k Y_k\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_k d_l\,\mathrm{Cov}(X_k, Y_l).$$

   In particular,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\mathrm{Var}(X_i) + \sum_{i\neq j}\mathrm{Cov}\left(X_i, X_j\right).\end{aligned}$$

Proof. For (3), we have

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

For (5), we have

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

Recall that if $X_1, .., X_n$ are discrete random variables, then they are independent iff for all $x_1, ..., x_n$,

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdot ... \cdot \mathbb{P}(X_n = x_n).$$

Proposition 3.25

If X_1, X_2, X_3 are independent, then X_1 is independent of X_2.

Proof. We need to show that

$$\forall x_1, x_2, \quad \mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2).$$

We have

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2) &= \sum_{x_3} \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ &= \sum_{x_3} \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\mathbb{P}(X_3 = x_3) \\ &= \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\underbrace{\sum_{x_3} \mathbb{P}(X_3 = x_3)}_{1}.\end{aligned}$$

Lemma 3.26

Let X and Y be 2 independent random variables, and $f, g : \mathbb{R} \to \mathbb{R}_+$. Then

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

Proof. Let $Z = (X, Y)$ and define $h(Z) = f(X)g(Y)$. Then

$$\begin{aligned}\mathbb{E}[h(\boldsymbol{Z})] &= \sum_{x,y} h(x,y)\mathbb{P}(\boldsymbol{Z} = (x,y)) \\ &= \sum_{x,y} f(x)g(y)\mathbb{P}(X = x, Y = y) \\ &= \sum_{x,y} f(x)g(y)\mathbb{P}(X = x)\mathbb{P}(Y = y) \\ &= \left(\sum_{x} f(x)\mathbb{P}(X = x)\right)\left(\sum_{y} g(y)\mathbb{P}(Y = y)\right) \\ &= \mathbb{E}[f(X)]\mathbb{E}[g(Y)].\end{aligned}$$

Lemma 3.27

Let X and Y be 2 independent random variables. Then

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

Important. The converse of the above lemma is not true.

Example 3.28

Let $X_1, X_2, X_3 \sim \mathrm{Ber}\left(\frac{1}{2}\right)$ be independent random variables. Let

$$\begin{aligned} Y_1 &= 2X_1 - 1, \quad & Y_2 &= 2X_2 - 1 \\ Z_1 &= Y_1 X_3, & Z_2 &= Y_2 X_3. \end{aligned}$$

Then $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 0$. Moreover,

$$\mathbb{E}[Z_1] = \mathbb{E}[Y_1 X_3] = \mathbb{E}[Y_1]\mathbb{E}[X_3] = 0, \quad \mathbb{E}[Z_2] = \mathbb{E}[Y_2 X_3] = \mathbb{E}[Y_2]\mathbb{E}[X_3] = 0.$$

So,

$$\begin{aligned}\mathrm{Cov}(Z_1, Z_2) &= \mathbb{E}[Z_1 Z_2] - \mathbb{E}[Z_1]\mathbb{E}[Z_2] \\ &= \mathbb{E}\left[Y_1 Y_2 X_3^2\right] - 0 \\ &= \mathbb{E}[Y_1 Y_2]\mathbb{E}\left[X_3^2\right] \\ &= \mathbb{E}[Y_1 Y_2]\mathbb{E}[X_3] \\ &= 0.\end{aligned}$$

However,

$$\begin{aligned}\mathbb{P}(Z_1 = 0, Z_2 = 0) &= \mathbb{P}(X_3 = 0) + \mathbb{P}(X_3 = 1, Y_1 = 0) + \mathbb{P}(X_3 = 1, Y_2 = 0) \\ &= \frac{1}{2} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{3}{4} \neq \mathbb{P}(Z_1 = 0)\mathbb{P}(Z_2 = 0) = \frac{1}{4}.\end{aligned}$$

Hence Z_1 and Z_2 are not independent, even though $\mathrm{Cov}(Z_1, Z_2) = 0$.

Lecture 9 · 2026-02-11

Corollary 3.29

Let $X_1, X_2, \ldots, X_n$ be independent random variables. Then

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

Example 3.30

Consider n random variables $S_n \sim \operatorname{Bin}(n,p)$. Then we have $\operatorname{Var}(S_n) = np(1-p)$ as seen before.

We also know that $S_n = X_1 + \ldots + X_n$ where (X_i) are independent and $X_i \sim \operatorname{Ber}(p)$.

Hence, by the above corollary, we have

$$\begin{aligned}\operatorname{Var}(S_n) &= \sum_{i=1}^{n} \operatorname{Var}(X_i) \\ &= \sum_{i=1}^{n} p(1-p) \\ &= np(1-p).\end{aligned}$$

3.3 Inequalities

Proposition 3.31 (Markov's Inequality)

Let X be a non-negative random variable and $a > 0$. Then

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

Proof. Observe that

$$X \geqslant a \cdot \mathbb{1}(X \geqslant a).$$

Taking expectation on both sides, we get

$$\mathbb{E}[X] \geqslant a\mathbb{E}[\mathbb{1}(X \geqslant a)] = a\mathbb{P}(X \geqslant a).$$

Proposition 3.32 (Chebyshev's Inequality)

Let X be a random variable and $a > 0$. Then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) \leqslant \frac{\operatorname{Var}(X)}{a^2}.$$

Proof. Note that

$$\begin{aligned}\mathbb{P}(|X-\mathbb{E}[X]| \geqslant a) &= \mathbb{P}\left(|X-\mathbb{E}[X]|^2 \geqslant a^2\right) \\ &\leqslant \frac{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]}{a^2} \quad \text{by Markov's} \\ &= \frac{\mathrm{Var}(X)}{a^2}.\end{aligned}$$

Proposition 3.33 (Cauchy-Schwarz Inequality)

Let X and Y be random variables. Then

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

Proof. Assume that $\mathbb{E}\left[X^2\right] < \infty$ and $\mathbb{E}\left[Y^2\right] < \infty$, otherwise there is nothing to prove. Then

$$|XY| \leqslant \frac{1}{2}\left(X^2+Y^2\right) \Rightarrow \mathbb{E}[|XY|] \leqslant \infty.$$

Assume that $\mathbb{E}\left[X^2\right] > 0$ and $\mathbb{E}\left[Y^2\right] > 0$, otherwise this is the trivial case. Assume WLOG that X and Y are non-negative.

Let $t \in \mathbb{R}$ and consider $(X-tY)^2 \geqslant 0$. Then

$$\begin{aligned}\mathbb{E}\left[(X-tY)^2\right] &\geqslant 0 \\ \underbrace{\mathbb{E}\left[X^2\right] - 2t\mathbb{E}[XY] + t^2\mathbb{E}\left[Y^2\right]}_{f(t)} &\geqslant 0.\end{aligned}$$

Then

$$f'(t) = -2\mathbb{E}[XY] + 2t\mathbb{E}\left[Y^2\right].$$

This function is minimised at $f(t_*)$ where $t_* = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$. Hence, using the fact that $f(t_*) \geqslant 0$, we get

$$\begin{aligned}&\mathbb{E}\left[X^2\right] - 2t_*\mathbb{E}[XY] + t_*^2\mathbb{E}\left[Y^2\right] \geqslant 0 \\ \Rightarrow\ &\mathbb{E}\left[X^2\right] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} \geqslant 0 \\ \Rightarrow\ &\mathbb{E}[|XY|] \leqslant \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.\end{aligned}$$

Remark. The equality holds when $f(t_*) = 0$, so

$$\mathbb{E}\left[(X-t_*Y)^2\right] = 0 \Rightarrow \mathbb{P}(X-t_*Y=0) = 1 \Rightarrow \mathbb{P}(X=t_*Y) = 1.$$

Definition 3.34 (Convex Function)

A function $f : \mathbb{R} \to \mathbb{R}$ is called **convex** if $\forall x, y \in \mathbb{R}$, $\forall t \in [0,1]$, we have

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y).$$

Proposition 3.35 (Jensen's Inequality)

Let X be a random variable and $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)].$$

Proof. We first need an additional claim.

> ***Claim.*** Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then f is the supremum of all the lines below it, i.e. for all $m \in \mathbb{R}$, $\exists a, b \in \mathbb{R}$ such that
>
> $$f(x) \geqslant ax + b \quad \text{for all } x \in \mathbb{R},$$
> $$f(m) = am + b.$$
>
> ***Proof.*** Let $m \in \mathbb{R}$ and $x < m < y$. Then for some $t \in (0,1)$, we have
>
> $$m = tx + (1-t)y \Leftrightarrow t(m-x) = (1-t)(y-m).$$
>
> By the convexity of f, we have
>
> $$f(m) = f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$$
> $$\Rightarrow t(f(m) - f(x)) \leqslant (1-t)(f(y) - f(m)).$$
>
> Hence, using the fact that $t(m-x) = (1-t)(y-m)$, we get
>
> $$\frac{f(m) - f(x)}{m - x} \leqslant \frac{f(y) - f(m)}{y - m}.$$
>
> Let $a = \sup_{x<m} \frac{f(m)-f(x)}{m-x}$. Then $\forall x < m < y$,
>
> $$\frac{f(m) - f(x)}{m - x} \leqslant a \leqslant \frac{f(y) - f(m)}{y - m}.$$
>
> So, this gives that $\forall z$,
>
> $$f(z) \geqslant a(z - m) + f(m).$$
>
> Hence, we can take $b = f(m) - am$ and get the desired result.

Let $m = \mathbb{E}[X]$. By the claim, $\exists a, b \in \mathbb{R}$ such that

$$f(x) \geqslant ax + b \quad \text{for all } x \in \mathbb{R},$$
$$f(m) = am + b.$$

Then we have $f(X) \geqslant aX + b$. Taking expectations on both sides, we get

$$\mathbb{E}[f(X)] \geqslant a\mathbb{E}[X] + b = am + b = f(m) = f(\mathbb{E}[X]).$$

Remark. For the equality case, let $f : \mathbb{R} \to \mathbb{R}$ be convex with the extra property that for $m = \mathbb{E}[X]$, $\exists a, b \in \mathbb{R}$, such that $f(x) > ax + b$ for all $x \in \mathbb{R} \setminus \{m\}$ and $f(m) = am + b$.

We wish to find a condition of X such that we have

$$\mathbb{E}[f(X)] = f[\mathbb{E}[X]].$$

We have $f(X) \geqslant aX + b \Rightarrow f(X) - (aX + b) \geqslant 0$. Taking expectations gives

$$\mathbb{E}[f(X)] \geqslant a\mathbb{E}[X] + b = am + b = f(m) = f(\mathbb{E}[X]).$$

Hence, the equality forces $\mathbb{E}[f(X) - (aX + b)] = 0$.

Since $f(X) - (aX + b) \geqslant 0$, we must have $\mathbb{P}(f(X) - (aX + b) = 0) = 1$. By the extra property of f, this is equivalent to $\mathbb{P}(X = m) = 1$.

Lecture 10 · 2026-02-13

Proposition 3.36 (Am-GM Inequality)

Let $x_1, x_2, \ldots, x_n \in \mathbb{R}_+$, then

$$\frac{1}{n}\sum_{k=1}^{n} x_k \geqslant \left(\prod_{k=1}^{n} x_k\right)^{\frac{1}{n}}.$$

Proof.

Claim. Let f be a convex function. Then $\forall x_1, \ldots, x_n \in \mathbb{R}$, we have

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k) \geqslant f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right).$$

Proof. Let X be a random variable with $\mathbb{P}(X = x_i) = \frac{1}{n}$ for all $i = 1, \ldots, n$. Then

$$\mathbb{E}[f(X)] = \frac{1}{n}\sum_{k=1}^{n} f(x_k) \geqslant f(\mathbb{E}[X]) = f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right).$$

Let $f(x) = -\log x$ with $x > 0$. This is a convex function. Hence, by the above claim, we have

$$\frac{1}{n}\sum_{k=1}^{n} -\log x_k \geqslant -\log\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right).$$

Rearranging gives the desired result.

3.4 Multiple Discrete Random Variables

3.4.1 Joint Distribution and Conditional Distribution

Recall that if X is a discrete random variable, then the distribution of X is given by $(\mathbb{P}(X = x))_x$.

Definition 3.37 (Joint Distrbution and Marginal Distribution)

Let $X_1, X_2, ..., X_n$ be discrete random variables. Their **joint distribution** is defined to be

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n) \quad \text{for all } x_1, ..., x_n.$$

In particular,

$$\mathbb{P}(X_1 = x_1) = \sum_{x_2,...,x_n} \mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

where $\mathbb{P}(X_1 = x_1)$ is called the **marginal distribution** of X_1.

Definition 3.38 (Conditional Distribution)

Let X and Y be two discrete random variables. The conditional distribution of X given $Y = y$ is defined by

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

By the law of total probability, we have

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

3.4.2 Distribution of the Sum of Random Variables

Let X and Y be two independent random variables. We wish to find the distribution of $X + Y$.

$$\begin{aligned}\mathbb{P}(X + Y = z) &= \sum_y \mathbb{P}(X = z - y, Y = y) \\ &= \underbrace{\sum_y \mathbb{P}(X = z - y)\mathbb{P}(Y = y).}_{\text{convolution of the 2 distributions}}\end{aligned}$$

Example 3.39 (Independent Poisson Random Variables)

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ with $X \perp Y$ Then

$$\begin{aligned}\mathbb{P}(X + Y = n) &= \sum_{k=0}^{n} \mathbb{P}(X = n - k)\mathbb{P}(Y = k) \\ &= \sum_{k=0}^{n} e^{-\lambda} \frac{\lambda^{n-k}}{(n-k)!} e^{-\mu} \frac{\mu^k}{k!} \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda^{n-k} \mu^k \\ &= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}.\end{aligned}$$

So $\mathbb{P}(X + Y = n) = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}$. Hence $X + Y \sim \text{Poi}(\lambda + \mu)$

3.4.3 Conditional Expectation

Recall that for $\mathbb{P}(B) > 0$,

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

Definition 3.40 (Conditional Expectation)

Let X be a discrete random variable and $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The **conditional expectation** of X given and event B is defined by

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X \cdot \mathbb{1}(B)]}{\mathbb{P}(B)}.$$

Note that if $X = \mathbb{1}(A)$, we recover

$$\mathbb{E}[\mathbb{1}(A) \mid B] = \frac{\mathbb{E}[\mathbb{1}(A \cap B)]}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A \mid B).$$

Proposition 3.41 (Law of Total Expectation)

Let X be a discrete random variable and (Ω_n) a partition of Ω with $\mathbb{P}(\Omega_n) > 0$ for all n. Then

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \mid \Omega_n]\mathbb{P}(\Omega_n).$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) \\
&= \sum_n \sum_{\omega \in \Omega_n} X(\omega)\mathbb{P}(\{\omega\}) \\
&= \sum_n \mathbb{E}[X \cdot \mathbb{1}(\Omega_n)] \\
&= \sum_n \mathbb{E}[X \mid \Omega_n]\mathbb{P}(\Omega_n).
\end{aligned}$$

In particular, for the conditional probability of X given $Y = y$, we have

$$\begin{aligned}
\mathbb{E}[X \mid Y = y] &= \frac{\mathbb{E}[X \cdot \mathbb{1}(Y = y)]}{\mathbb{P}(Y = y)} \\
&= \sum_x x \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \\
&= \sum_x x\mathbb{P}(X = x \mid Y = y).
\end{aligned}$$

Let $g(y) = \mathbb{E}[X \mid Y = y]$ since this is a function of y. Consider what is meant by $\mathbb{E}[X|Y]$.

We will define $\mathbb{E}[X \mid Y]$ as a random variable, which is a function of Y.

Definition 3.42 (Conditional Expectation Given a Random Variable)

Let X and Y be discrete random variables. The **conditional expectation** of X given Y is defined by

$$\mathbb{E}[X \mid Y] = \sum_y \mathbb{E}[X \mid Y = y]\mathbb{1}(Y = y) = g(Y).$$

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

Remark. The above notation can be confusing. $\mathbb{E}[X \mid Y]$ is a random variable, which is a function of Y. In particular, if $Y(\omega) = y$, then $\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = y]$.

Bear in mind that random variables are functions. We have $X, Y : \Omega \to \mathbb{R}$, and so $g(Y)$ really means $g \circ Y$.

Example 3.43

Consider tossing a p-coin n times independently. For $i = 1, \ldots, n$, let

$$X_i = \mathbb{1}(i - \text{th toss is H})$$
$$Y_n = X_1 + \ldots + X_n$$

We wish to find $\mathbb{E}[X_1 \mid Y_n]$. Let

$$\begin{aligned} g(y) &= \mathbb{E}[X_1 \mid Y_n = y] \\ &= \frac{\mathbb{E}[X_1 \cdot \mathbb{1}(Y_n = y)]}{\mathbb{P}(Y_n = y)} \\ &= \frac{\mathbb{P}(X_1 = 1, Y_n = y)}{\mathbb{P}(Y_n = y)} \\ &= p \cdot \frac{\binom{n-1}{y-1} p^{y-1}(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{y}{n}. \end{aligned}$$

So, $\mathbb{E}[X_1 \mid Y_n] = g(Y_n) = \frac{Y_n}{n}$.

Proposition 3.44

Let X and Y be discrete random variables. Then for some constant $c \in \mathbb{R}$,

1. $\mathbb{E}[cX \mid Y] = c\mathbb{E}[X \mid Y]$.
2. $\mathbb{E}[c + X \mid Y] = c + \mathbb{E}[X \mid Y]$.

   In particular, $\mathbb{E}[c \mid Y] = c$.
3. $\mathbb{E}[\sum_{i=1}^n X_i \mid Y] = \sum_{i=1}^n \mathbb{E}[X_i \mid Y]$.

Lecture 11 · 2026-02-16

Proposition 3.45 (Tower Property)

Let X and Y be discrete random variables. Then

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

Proof. We have

$$\mathbb{E}[X \mid Y] = \sum_y \mathbb{E}[X \mid Y = y]\mathbb{1}(Y = y).$$

Taking expectation on both sides gives

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) \\ &= \mathbb{E}[X] \quad \text{by the law of total expectation.}\end{aligned}$$

Proposition 3.46

Let X and Y be independent discrete random variables. Then

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

Proof. We have

$$\begin{aligned}\mathbb{E}[X \mid Y] &= \sum_y \mathbb{E}[X \mid Y = y]\mathbb{1}(Y = y) \\ &= \sum_y \mathbb{1}(Y = y) \sum_x x\mathbb{P}(X = x \mid Y = y) \\ &= \sum_y \mathbb{1}(Y = y) \sum_x x\mathbb{P}(X = x) \quad \text{by independence} \\ &= \sum_y \mathbb{1}(Y = y)\mathbb{E}[X] \\ &= \mathbb{E}[X].\end{aligned}$$

Proposition 3.47

Let Y, Z be independent discrete random variables. Then for any random variable X,

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

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

> ***Claim.*** $g(Y)$ is independent of Z.
>
> ***Proof.*** We need to show that

$$\forall w, z \in \mathbb{R}, \quad \mathbb{P}(g(Y) = w, Z = z) = \mathbb{P}(g(Y) = w)\mathbb{P}(Z = z).$$

We have

$$\begin{aligned}\mathbb{P}(g(Y) = w, Z = z) &= \sum_{y \in g^{-1}(\{w\})} \mathbb{P}(Y = y, Z = z) \\ &= \sum_{y \in g^{-1}(\{w\})} \mathbb{P}(Y = y)\mathbb{P}(Z = z) \quad \text{by independence} \\ &= \mathbb{P}(g(Y) = w)\mathbb{P}(Z = z).\end{aligned}$$

Then, by Proposition 3.47, we have $\mathbb{E}[g(Y) \mid Z] = \mathbb{E}[g(Y)]$.

By Proposition 3.46, we have $\mathbb{E}[g(Y)] = \mathbb{E}[X]$. Hence, $\mathbb{E}[\mathbb{E}[X \mid Y] \mid Z] = \mathbb{E}[X]$.

Proposition 3.48

Let X, Y be two random variables and $h : \mathbb{R} \to \mathbb{R}$. Then

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

Proof. We have

$$\begin{aligned}\mathbb{E}[h(Y) \cdot X \mid Y] &= \sum_y \mathbb{1}(Y = y)\mathbb{E}[h(Y) \cdot X \mid Y = y] \\ &= \sum_y h(y)\mathbb{1}(Y = y)\mathbb{E}[X \mid Y = y] \\ &= h(Y)\mathbb{E}[X \mid Y].\end{aligned}$$

Corollary 3.49

Let X, Y be two random variables. Then

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

Example 3.50

Consider tossing a p-coin n times independently. For $i = 1, \ldots, n$, let

$$\begin{aligned}X_i &= \mathbb{1}(i - \text{th toss is H}) \\ Y_n &= X_1 + \ldots + X_n\end{aligned}$$

We wish to find $\mathbb{E}[X_1 \mid Y_n]$. By symmetry, $\mathbb{E}[X_i \mid Y_n] = \mathbb{E}[X_1 \mid Y_n]$ for all i. Hence,

$$\begin{aligned} Y_n = \mathbb{E}[Y_n \mid Y_n] &= \mathbb{E}[X_1 + \ldots + X_n \mid Y_n] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i \mid Y_n] \\ &= n\mathbb{E}[X_1 \mid Y_n]. \end{aligned}$$

So, $\mathbb{E}[X_1 \mid Y_n] = \frac{Y_n}{n}$.

3.5 Random Walks

Definition 3.51 (Random Process / Stochastic Process)

A **random process** or **stochastic process** is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$.

Definition 3.52 (Random Walk)

A **random walk** is a random process that can be expressed as

$$X_n = x + Y_1 + Y_2 + \ldots + Y_n$$

where x is a deterministic constant and $(Y_n)_{n\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) random variables.

The steps of the random walk are the random variables Y_n.

3.5.1 Simple Random Walk

Definition 3.53 (Simple Random Walk)

A **simple random walk** is a random walk satisfying $\mathbb{P}(Y_i = 1) = p = 1 - \mathbb{P}(Y_i = -1)$.

If $p = q = 1 - p = \frac{1}{2}$, then we call it a simple **symmetric** random walk.

Example 3.54 (Gambler's Ruin)

Consider (X_n) as a simple random walk, which is the fortune of a gambler who starts with £x at time 0 and at every time step, he wins £1 with probability p and loses £1 with probability $q = 1 - p$.The game ends if he reaches 0 or if he reaches £a, whichever comes first.

> ***Notation.*** We will denote $\mathbb{P}_x(A) = \mathbb{P}(A \mid X_0 = x)$, and $X = (X_n)$.

Let $h(x) = \mathbb{P}_x((X_n)$ reaches a before reaching $0)$. Then $h(0) = 0$ and $h(a) = 1$.

Then,

$$\begin{aligned} h(x) &= \mathbb{P}_x(X \text{ reaches } a \text{ before reaching } 0) \\ &= \mathbb{P}_x(X \text{ reaches } a \text{ before } 0, Y_1 = 1) + \mathbb{P}_x(X \text{ reaches } a \text{ before } 0, Y_1 = -1) \\ &= p\mathbb{P}_{x+1}(X \text{ reaches } a \text{ before } 0 \mid Y_1 = 1) + (1-p)\mathbb{P}_{x-1}(X \text{ reaches } a \text{ before } 0 \mid Y_1 = -1) \\ &= p \cdot h(x+1) + (1-p) \cdot h(x-1). \end{aligned}$$

So we can solve the following system of equations to find $h(x)$ for all x:

$$\begin{cases} h(x) = ph(x+1) + (1-p)h(x-1) \\ h(0) = 0 \\ h(a) = 1 \end{cases}$$

Lecture 12 · 2026-02-18

- For $p = q = \frac{1}{2}$, we get a simple symmetric random walk (SSRW), in which case we have

  $$\begin{cases} h(x) = \frac{1}{2}h(x+1) + \frac{1}{2}h(x-1) \\ h(0) = 0 \\ h(a) = 1 \end{cases}$$

  This leads to

  $$h(x+1) - h(x) = h(x) - h(x-1) = c.$$

  Hence,

  $$h(x) = \sum_{i=1}^{x} (h(i) - h(i-1)) = \sum_{i=1}^{x} c = cx.$$

  Considering boundary conditions, $h(a) = 1$ gives $c = \frac{1}{a}$. Hence, for a SSRW, we have

  $$h(x) = \frac{x}{a}.$$

- For $p \neq q$, we need to try a solution of the form λ^x for some λ. Then

  $$\lambda^x = p \cdot \lambda^{x+1} + q\lambda^{x-1} \Rightarrow \lambda = 1 \text{ or } \lambda = \frac{q}{p}.$$

  The general solution is of the form $h(x) = A + B\left(\frac{q}{p}\right)^x$ for some constants $A, B \in \mathbb{R}$. Using the boundary conditions, we get

  $$h(x) = \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}.$$

3.5.2 Time to Absorption

Let (X_n) be a simple random walk. We are interested in the time to absorption, which is the duration of the game. Let the time to absorption be denoted by T. Then

$$T = \min\{n \geqslant 0 : X_n \in \{0, a\}\}.$$

We would like to consider the expected time to absorption, i.e. $\mathbb{E}[T \mid X_0 = x]$.

Notation. We will denote $\mathbb{E}_x[T] = \mathbb{E}[T \mid X_0 = x] = k(x)$.

By the law of total expectation, we have

$$\begin{aligned} k(x) &= \mathbb{E}_x[T] \\ &= \mathbb{E}_x[T \mid Y_1 = 1]\mathbb{P}(Y_1 = 1) + \mathbb{E}_x[T \mid Y_1 = -1]\mathbb{P}(Y_1 = -1) \\ &= p\mathbb{E}_{x+1}[T+1] + q\mathbb{E}_{x-1}[T+1] \\ &= p(k(x+1)+1) + q(k(x-1)+1) \\ &= pk(x+1) + qk(x-1) + 1. \end{aligned}$$

Boundary conditions are $k(0) = k(a) = 0$.

- For $p = q = \frac{1}{2}$, try a solution of the form $k(x) = Ax^2$. Then

$$Ax^2 = \frac{1}{2}A(x+1)^2 + \frac{1}{2}A(x-1)^2 + 1 \Rightarrow A = -1.$$

  The general solution will be of the form $k(x) = -x^2 + Bx + C$. Using boundary conditions, we get

$$k(x) = x(a-x).$$

- For $p \neq q$, try Cx as a solution. Then $C = \frac{1}{q-p}$. Hence

$$k(x) = A + \frac{1}{q-p}x + B\left(\frac{q}{p}\right)^x.$$

  Solving this gives

$$k(x) = \frac{1}{q-p}x - \frac{a}{q-p}\frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}.$$

3.6 Probability Generating Functions

3.6.1 Introduction

Definition 3.55 (Probability Mass Function)

Let X be a discrete random variable. The **probability mass function** of X is defined by

$$p_r = \mathbb{P}(X = r).$$

Definition 3.56 (Probability Generating Function)

Let X be a discrete random variable taking values in $\mathbb{N}$. The **probability generating function** (PGF) of X is defined by

$$p(z) = \mathbb{E}\left[z^X\right] = \sum_{r=0}^{\infty} p_r z^r$$

where $|z| \leqslant 1$.

Note that $p(z)$ converges absolutely for $|z| \leqslant 1$ since $p_r \geqslant 0$ and $\sum_{r=0}^{\infty} p_r = 1$. Hence the radius of convergence of $p(z)$ is at least 1. Therefore, $p(z)$ is well-defined for $|z| \leqslant 1$.

Remark. In this section, we will only consider discrete random variables taking values in $\mathbb{N}$.

Theorem 3.57

The PGF of a random variable X uniquely determines the distribution of X.

Proof. Let (p_r) and (q_r) be 2 probability distributions with the same PGF, *i.e.* for all $|z| \leqslant 1$,

$$\sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} q_r z^r.$$

We need to show that $p_r = q_r$ for all r. We shall show this by induction. For $z \to 0$,

$$p_0 = q_0.$$

Assume that $p_r = q_r$ for all $r \leqslant n$. Then

$$\sum_{r=n+1}^{\infty} p_r z^r = \sum_{r=n+1}^{\infty} q_r z^r.$$

Dividing both sides by z^{n+1} and letting $z \to 0$ gives

$$p_{n+1} = q_{n+1}.$$

Example 3.58

Consider $X \sim \text{Bin}(n, p)$. Then

$$p(z) = \mathbb{E}\left[z^X\right] = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} z^r = (1 - p + pz)^n.$$

Remark. Suppose that $X_1, \ldots, X_n$ are independent random variables with PGFs

$$q_i(z) = \mathbb{E}\left[z^{X_i}\right].$$

Consider the PGF of $S_n = X_1 + \ldots + X_n$:

$$p(z) = \mathbb{E}\left[z^{X_1 + \ldots + X_n}\right] = \mathbb{E}\left[z^{X_1} z^{X_2} \ldots z^{X_n}\right] = \mathbb{E}\left[z^{X_1}\right]\mathbb{E}\left[z^{X_2}\right]\ldots\mathbb{E}\left[z^{X_n}\right] = \prod_{i=1}^{n} q_{i(z)}.$$

Example 3.59

Consider $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ with $X \perp Y$. Then

$$\mathbb{E}\left[z^{X+Y}\right] = \mathbb{E}\left[z^X\right]\mathbb{E}\left[z^Y\right] = (1 - p + pz)^n (1 - p + pz)^m = (1 - p + pz)^{n+m}.$$

Hence $X + Y \sim \text{Bin}(n + m, p)$.

Example 3.60

Consider $X \sim \text{Geo}(p)$. Then

$$\mathbb{E}\left[z^X\right] = \sum_{r=1}^{\infty} z^r (1-p)^{r-1} p = \frac{pz}{1-z(1-p)}.$$

Example 3.61

Consider $X \sim \text{Poi}(\lambda)$. Then

$$\mathbb{E}\left[z^X\right] = \sum_{r=0}^{\infty} z^r e^{-\lambda} \frac{\lambda^r}{r!} = e^{-\lambda} \sum_{r=0}^{\infty} \frac{(\lambda z)^r}{r!} = e^{-\lambda} e^{\lambda z} = e^{\lambda(z-1)}.$$

Example 3.62

Consider $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ with $X \perp Y$. Then

$$\mathbb{E}\left[z^{X+Y}\right] = \mathbb{E}\left[z^X\right]\mathbb{E}\left[z^Y\right] = e^{\lambda(z-1)} e^{\mu(z-1)} = e^{(\lambda+\mu)(z-1)}.$$

Hence, $X + Y \sim \text{Poi}(\lambda + \mu)$.

Theorem 3.63

Let $p(z)$ be the PGF of a random variable X. Then

$$\lim_{z \to 1^-} p'(z) = p'(1^-) = \mathbb{E}[X].$$

Proof.

- Assume that $\mathbb{E}[X] < \infty$. Take $0 < z < 1$. Then

$$p'(z) = \sum_{r=1}^{\infty} r p_r z^{r-1} \leqslant \sum_{r=1}^{\infty} r p_r = \mathbb{E}[X].$$

  So $p'(z)$ is increasing in z and bounded above by $\mathbb{E}[X]$. Hence, $\lim_{z \to 1^-} p'(z) \leqslant \mathbb{E}[X]$.

  Take $\varepsilon > 0$. There exists N such that

$$\sum_{r=1}^{N} r p_r \geqslant \mathbb{E}[X] - \varepsilon.$$

  Then,

$$\lim_{z \to 1^-} p'(z) \geqslant \lim_{z \to 1^-} \sum_{r=1}^{N} r p_r z^{r-1} = \sum_{r=1}^{N} r p_r \geqslant \mathbb{E}[X] - \varepsilon.$$

  Hence, $\lim_{z \to 1^-} p'(z) = \mathbb{E}[X]$.

- Assume that $\mathbb{E}[X] = \infty$. Then $\forall M > 0, \exists N$ such that

$$\sum_{r=1}^{N} r p_r \geqslant M.$$

Then,

$$\lim_{z \to 1^-} p'(z) \geqslant \lim_{z \to 1^-} \sum_{r=1}^{N} r p_r z^{r-1} = \sum_{r=1}^{N} r p_r \geqslant M.$$

Hence, $\lim_{z \to 1^-} p'(z) = \infty$.

Theorem 3.64

Let $p(z)$ be the PGF of a random variable X. Then

- $p''(1^-) = \mathbb{E}[X(X-1)]$.
- $\forall k \geqslant 0, p^{(k)}(1^-) = \mathbb{E}[X(X-1)...(X-k+1)]$.
- $\mathrm{Var}(X) = p''(1^-) + p'(1^-) - (p'(1^-))^2$.
- $\mathbb{P}(X = n) = \frac{\mathrm{d}^n}{\mathrm{d}z^n}\left(p(z) \cdot \frac{1}{n!}\right)\Big|_{z=0}$

Lecture 13 · 2026-02-20

3.6.2 Sum of a Random Number of Random Variables

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables and let N be an independent random variable taking values in $\mathbb{N}$.

Let

$$S_n = X_1 + X_2 + \ldots + X_n$$

$$S_n(\omega) = \sum_{i=1}^{n} X_{i(\omega)}$$

$$S_N(\omega) = \sum_{i=1}^{N(\omega)} X_{i(\omega)}.$$

Lemma 3.65

Let $q(z) = \mathbb{E}\left[z^N\right]$ be the PGF of N and $p(z) = \mathbb{E}\left[z^{X_1}\right]$ be the PGF of X_1.

Then the PGF of S_N is given by

$$r(z) = \mathbb{E}\left[z^{S_N}\right] = q(p(z)).$$

Proof.

Proof 1 We have

$$\begin{aligned} r(z) = \mathbb{E}\left[z^{S_N}\right] &= \sum_{n=0}^{\infty} \mathbb{E}\left[z^{S_n} \cdot \mathbb{1}(N=n)\right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}\left[z^{X_1+\ldots+X_n} \cdot \mathbb{1}(N=n)\right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}\left[z^{X_1} z^{X_2} \ldots z^{X_n}\right] \cdot \mathbb{P}(N=n) \\ &= \sum_{n=0}^{\infty} \left(\mathbb{E}\left[z^{X_1}\right]\right)^n \cdot \mathbb{P}(N=n) \\ &= \sum_{n=0}^{\infty} (p(z))^n \cdot \mathbb{P}(N=n) \\ &= q(p(z)) \quad \text{by definition of PGF.} \end{aligned}$$

Proof 2 We have

$$r(z) = \mathbb{E}\left[z^{S_N}\right] = \mathbb{E}\left[\mathbb{E}\left[z^{S_N} \;\middle|\; N\right]\right].$$

Note that

$$\begin{aligned} \mathbb{E}\left[z^{S_N} \;\middle|\; N\right](n) &= \mathbb{E}\left[z^{S_N} \;\middle|\; N=n\right] \\ &= \mathbb{E}\left[z^{S_n} \;\middle|\; N=n\right] \\ &= \mathbb{E}\left[z^{S_n}\right] \quad \text{by independence} \\ &= (p(z))^n. \end{aligned}$$

Hence, we can write $\mathbb{E}\left[z^{S_N} \;\middle|\; N\right] = (p(z))^N$. Taking expectation on both sides gives

$$r(z) = \mathbb{E}\left[(p(z))^N\right] = q(p(z)).$$

We have

$$\begin{aligned} \mathbb{E}[S_N] &= r'(1^-) = q'(p(1^-))p'(1^-) = \mathbb{E}[N]\mathbb{E}[X_1] \\ \operatorname{Var}(S_N) &= \mathbb{E}[N]\operatorname{Var}(X_1) + \operatorname{Var}(N)(\mathbb{E}[X_1])^2. \end{aligned}$$

3.7 Branching Processes

Let $X_0 = 1$, and X_1 be the distribution of the number of offspring of the individual in the first generation. [X_n is the number of individuals in the n-th generation.]

$$\mathbb{P}(X_1 = k) = g_k \quad \forall k \in \mathbb{Z}_{\geqslant 0}.$$

Each individual produces an independent number of offspring with the same distribution as X_1.

Let $(Y_{k,n})_{k\geqslant 1, n\geqslant 0}$ be a family of independent random variables such that $Y_{k,n}$ has the same distribution as X_1 for all k, n. Then we can write

$$Y_{k,n} = \text{number of offspring the } k\text{-th individual in the } n\text{-th generation produces,}$$

$$X_{n+1} = \begin{cases} 0 & \text{if } X_n = 0 \\ \sum_{k=1}^{X_n} Y_{k,n} & \text{if } X_n > 0. \end{cases}$$

We are interested in the probability of extinction.

3.7.1 Generating Functions of Branching Processes

Theorem 3.66

Let (X_i) be a branching process with offspring distribution X_1. Then,

$$\mathbb{E}[X_n] = (\mathbb{E}[X_1])^n.$$

Proof. We will show this by induction. For $n = 0$, we have $\mathbb{E}[X_0] = 1 = (\mathbb{E}[X_1])^0$. Assume that $\mathbb{E}[X_n] = (\mathbb{E}[X_1])^n$. Then

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[\mathbb{E}[X_{n+1} \mid X_n]].$$

Note that

$$\begin{aligned} \mathbb{E}[X_{n+1} \mid X_n = m] &= \mathbb{E}[Y_{1,n} + \ldots + Y_{m,n} \mid X_n = m] \\ &= \mathbb{E}[Y_{1,n} + \ldots + Y_{m,n}] \quad \text{by independence} \\ &= m\mathbb{E}[X_1]. \end{aligned}$$

Thus,

$$\mathbb{E}[X_{n+1} \mid X_n] = X_n \mathbb{E}[X_1].$$

Taking expectation on both sides gives

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n]\mathbb{E}[X_1] = (\mathbb{E}[X_1])^{n+1}.$$

Theorem 3.67

Let $G(z) = \mathbb{E}\left[z^{X_1}\right]$ and $G_n(z) = \mathbb{E}\left[z^{X_n}\right]$. Then

$$G_{n+1}(z) = G\left(G_{n(z)}\right) = G_n(G(z)) = G \circ G \circ \cdots \circ G(z).$$

Proof. We have

$$\begin{aligned} G_{n+1}(z) &= \mathbb{E}\left[z^{X_{n+1}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[z^{X_{n+1}} \;\middle|\; X_n\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[z^{Y_{1,n}+\ldots+Y_{X_n,n}} \;\middle|\; X_n\right]\right]. \end{aligned}$$

Note that,

$$\begin{aligned}\mathbb{E}\left[z^{Y_{1,n}+\ldots+Y_{X_n,n}} \mid X_n = m\right] &= \mathbb{E}\left[z^{Y_{1,n}+\ldots+Y_{m,n}}\right] \\ &= \mathbb{E}\left[z^{Y_{1,n}}\right]\ldots\mathbb{E}\left[z^{Y_{m,n}}\right] \quad \text{by independence} \\ &= (G(z))^m.\end{aligned}$$

Hence,

$$G_{n+1}(z) = \mathbb{E}\left[(G(z))^{X_n}\right] = G_n(G(z)).$$

3.7.2 Extinction Probability

Let

$$q = \mathbb{P}(X_n = 0 \text{ for some } n \geqslant 0)$$

and $q_n = \mathbb{P}(X_n = 0)$.

Since $\{X_n = 0\} \subseteq \{X_{n+1} = 0\}$, (q_n) is an increasing sequence that converges to q, by the continuity of probability measures.

[Recall that $A_n \subseteq A_{n+1}$ for all n implies that $\mathbb{P}(A_n) \to \mathbb{P}(\bigcup_n A_n)$.]

Proposition 3.68

Let $G(z)$ be the PGF of X_1. Then $G(z) = \mathbb{E}\left[z^{X_1}\right]$, and

$$q_{n+1} = G(q_n), \quad q = G(q).$$

Proof. If we were given that $q_{n+1} = G(q_n)$, then since G is continuous as $q_n \to q$, we have $q = G(q)$. So we just need to show that $q_{n+1} = G(q_n)$.

Proof 1 We have

$$\begin{aligned}q_{n+1} = \mathbb{P}(X_{n+1} = 0) &= G_{n+1}(0) \\ &= G(G_n(0)) \\ &= G(q_n).\end{aligned}$$

Proof 2 Conditioning on $X_1 = m$, let $X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(m)}$ be independent and identically distributed branching processes with the same offspring distribution as X_1. Then

$$X_{n+1} = \sum_{i=1}^{m} X_n^{(i)}.$$

Hence,

$$\begin{aligned} q_{n+1} &= \mathbb{P}(X_{n+1} = 0) \\ &= \sum_m \mathbb{P}(X_{n+1} = 0 \mid X_1 = m)\mathbb{P}(X_1 = m) \\ &= \sum_m \mathbb{P}\left(\sum_{i=1}^{m} X_n^{(i)} = 0 \;\middle|\; X_1 = m\right)\mathbb{P}(X_1 = m) \\ &= \sum_m \mathbb{P}\left(X_n^{(1)} = \ldots = X_n^{(m)} = 0\right)\mathbb{P}(X_1 = m) \\ &= \sum_m (q_n)^m \mathbb{P}(X_1 = m) \\ &= G(q_n). \end{aligned}$$

Lecture 14 · 2026-02-23

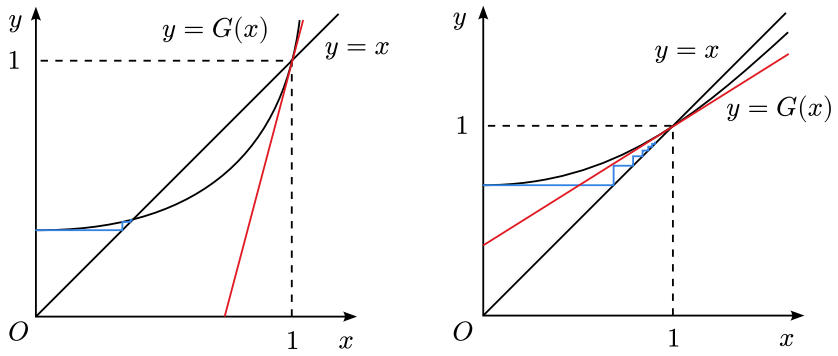
Note that in Theorem 3.66, we have

$$\mathbb{E}[X_n] = (\mathbb{E}[X_1])^n.$$

So, in order to evaluate q, consider the following cases:

- If $\mathbb{E}[X_1] < 1$, then $\mathbb{E}[X_n] \to 0$ as $n \to \infty$, so we expect that $q = 1$.
- If $\mathbb{E}[X_1] > 1$, then $\mathbb{E}[X_n] \to \infty$ as $n \to \infty$, so we expect that $q < 1$.
- If $\mathbb{E}[X_1] = 1$, then $\mathbb{E}[X_n] = 1$ for all n, which does not give us an intuition on q.

Note that the relation $q = G(q)$ always has a solution at $q = 1$, as shown in the following graphs:



Note that the gradient of the tangent at $x = 1$ is $G'(1) = \mathbb{E}[X_1]$. Hence, the first two cases can be justified.

Theorem 3.69

Let (X_n) be a branching process with offspring distribution X_1. Assume that $\mathbb{P}(X_1 = 1) < 1$.

Then the extinction probability q is the minimal non-negative solution to the equation $G(x) = x$.

Moreover, $q < 1$ iff $\mathbb{E}[X_1] > 1$.

Proof. We know that $q = G(q)$ in Proposition 3.68.

Let t be the smallest non-negative solution to $G(x) = x$. We will show that $q = t$.

We will prove by induction that $q_n \leqslant t$ for all n, which will imply that $q \leqslant t$ as $n \to \infty$, but since t is the smallest non-negative solution to $G(x) = x$, we have $t \leqslant q$. Hence, $q = t$.

For $n = 0$, we have $q_0 = \mathbb{P}(X_0 = 0) = 0 \leqslant t$. Assume that $q_n \leqslant t$. Then

$$q_{n+1} = G(q_n) \leqslant G(t) = t$$

since G is increasing in $[0, 1]$. Hence, $q_{n+1} \leqslant t$.

Now, we will show that $q < 1$ iff $\mathbb{E}[X_1] > 1$.

Assume that $\mathbb{P}(X_1 \leqslant 1) = g_0 + g_1 = 1$, then $\mathbb{P}(X_1 \leqslant 1) = 1$, and then

$$\mathbb{E}[X_1] = g_1.$$

Then in this case, $G(z) = g_0 + g_1 z = 1 - \mathbb{E}[X_1] + \mathbb{E}[X_1]z$. Then,

$$G(z) = z \Rightarrow (1 - \mathbb{E}[X_1]) \cdot z = 1 - \mathbb{E}[X_1].$$

Because $\mathbb{E}[X_1] = g_1 < 1$, we have $1 - \mathbb{E}[X_1] \neq 0$, and so $z = 1$.

Now assume that $g_1 < 1$ and $g_0 + g_1 < 1$, we shall show that $q < 1$ iff $\mathbb{E}[X_1] > 1$. Define

$$H(z) = G(z) - z = \sum_{r=0}^{\infty} g_r z^r - z.$$

Then $H(1) = 0$. We shall first show that H can have at most one more root in $[0, 1)$. We have

$$H''(z) = \sum_{r=0}^{\infty} r(r-q) g_r z^{r-2} > 0 \text{ in } (0, 1)$$

because $g_0 + g_1 < 1$ implies that there exists $r \geqslant 2$ such that $g_r > 0$.

Hence $H'(z)$ is strictly increasing in $(0, 1)$. Therefore, H can have at most one more root other than 1, due to Rolle's theorem. [If not, then $\exists z_1, z_2 < 1$ such that $H(z_1) = H(z_2) = 0$. Then by Rolle's theorem, there exists $z_3 \in (z_1, z_2)$ such that $H'(z_3) = 0$. Also, $\exists z_4 \in (z_2, 1)$ such that $H'(z_4) = 0$. This contradicts the fact that $H'(z)$ is strictly increasing in $(0, 1)$.]

We therefore have two cases:

- H has no other root other than 1. Then $q = 1$. We need to show that $\mathbb{E}[X_1] \leqslant 1$. We have

  $$H'(1^-) = G'(1^-) - 1 = \mathbb{E}[X_1] - 1 \leqslant 0.$$

  This is because $H(0) = G(0) = g_0 \geqslant 0$ and $H(1) = 0$. So

  $$H(z) \geqslant 0 \quad \forall z \in [0, 1].$$

  Hence

  $$H'(1^-) = \lim_{z \to 1^-} \frac{H(z) - H(1)}{z - 1} \leqslant 0.$$

  So $H'(1^-) \leqslant 0$, and hence $\mathbb{E}[X_1] \leqslant 1$.

- H has another root $r < 1$. So r has to be the extinction probability q. We need to show that $\mathbb{E}[X_1] > 1$. We have

$$H(r) = 0, \quad H(1) = 0.$$

By Rolle's theorem, there exists $z \in (r, 1)$ such that $H'(z) = 0$. We also have

$$H'(1^-) = \mathbb{E}[X_1] - 1.$$

Since H' is strictly increasing in $(0, 1)$, we have $H'(1^-) > H'(z) = 0$. Hence, $\mathbb{E}[X_1] > 1$.

4 Continuous Random Variables

4.1 Probability Distribution Function

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Recall that a random variable is a function

$$X : \omega \to \mathbb{R}$$

where $\forall x \in \mathbb{R}$, the set $X^{-1}(\{x\}) = \{X \leqslant x\} = \{\omega \in \Omega : X(\omega) = x\}$ is an event in $\mathcal{F}$.

Recall Definition of Probability Distribution Function 3.3. Let $F(x)$ be the probability distribution function of X. We have the following properties of F:

Proposition 4.1 (Properties of Probability Distribution Function)

Let F be the probability distribution function of a random variable X. Then, we have the following properties:

1. F is increasing, i.e. $F(x) \leqslant F(y)$ for all $x \leqslant y$.
2. For all $a, b \in \mathbb{R}$ with $a < b$, we have
$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a).$$
3. F is right-continuous and F always has left limits, i.e. for all $x \in \mathbb{R}$, we have
$$F(x^-) = \lim_{y \to x^-} F(y) \leqslant F(x)$$
$$F(x^+) = \lim_{y \to x^+} F(y) = F(x).$$
4. $F(x^-) = \mathbb{P}(X < x)$.
5. $F(-\infty) = 0$ and $F(\infty) = 1$.

Lecture 15 · 2026-02-25

Proof.

For (2), we have

$$\begin{aligned}
\mathbb{P}(a < X \leqslant b) &= \mathbb{P}(X \leqslant b, X > a) \\
&= \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant b, X \leqslant a) && \text{by the law of total probability} \\
&= F(b) - \mathbb{P}(X \leqslant a) && \text{since } a < b \\
&= F(b) - F(a).
\end{aligned}$$

For (3), let (x_n) be a decreasing sequence with $x_n \to x$. Then, we want to show that $F(x_n) \to F(x)$. Define

$$A_n = \{x < X \leqslant x_n\}$$

so

$$\mathbb{P}(A_n) = F(x_n) - F(x).$$

Note that $A_{n+1} \subseteq A_n$, and hence $\mathbb{P}(A_n) \to \mathbb{P}(\bigcap_n A_n)$ by the continuity of probability measures.

Now, we have $\bigcap_n A_n = \{x < X \leqslant x\} = \varnothing$, so $\mathbb{P}(\bigcap_n A_n) = 0$. Therefore, $F(x_n) \to F(x)$.

Since F is increasing, left limits always exist, and so $F(x^-) = \lim_{y \to x^-} F(y) \leqslant F(x)$.

For (4), consider

$$F(x^-) = \lim_{n \to \infty} F\left(x - \frac{1}{n}\right).$$

Define $B_n = \left\{X \leqslant x - \frac{1}{n}\right\}$. Then we have $B_n \subseteq B_{n+1}$. Hence

$$\mathbb{P}(B_n) \to \mathbb{P}\left(\bigcup_n B_n\right) = \mathbb{P}(X < x) = F(x^-).$$

Definition 4.2 (Continuous Random Variable)

A random variable X is said to be **continuous** if its probability distribution function F is continuous. Equivalently, for all $x \in \mathbb{R}$,

$$\begin{aligned} & F(x^-) = F(x) \\ \Leftrightarrow \quad & \mathbb{P}(X < x) = \mathbb{P}(X \leqslant x) \\ \Leftrightarrow \quad & \mathbb{P}(X = x) = 0. \end{aligned}$$

For the purpose of this course, we will only consider continuous random variables whose PDF is differentiable. These are also known as **absolutely continuous** random variables.

Definition 4.3 (Probability Density Function)

Let X be a continuous random variable with probability distribution function F. If F is differentiable, then the **probability density function** f of X is defined as

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

Proposition 4.4 (Properties of Probability Density Function)

Let X be a continuous random variable with probability density function f. Then, we have the following properties:

1. $f(x) \geqslant 0$ for all $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.
3. $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ for all $x \in \mathbb{R}$. More generally, for every $A \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in A) = \int_A f(x)\,\mathrm{d}x.$$

[If X is discrete, then $\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x)$, so the above formula can be seen as a continuous analogue of the discrete case.]

Intuitively, f can be thought as being proportional to the probability that X takes values around x. Mathematically, for Δx small, we have

$$\mathbb{P}(x < X \leqslant x + \Delta x) = \int_x^{x+\Delta x} f(t)\,\mathrm{d}t \approx f(x)\Delta x.$$

4.2 Uniform Distribution

Definition 4.5 (Uniform Distribution)

A **uniform distribution** on $[a, b]$ is a continuous random variable X, denoted $U[a, b]$, with probability density function

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

> ***Proof of density validity.*** We have
>
> $$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= \int_a^b \frac{1}{b-a}\,\mathrm{d}x \\ &= \frac{1}{b-a}\int_a^b \mathrm{d}x \\ &= \frac{1}{b-a}(b-a) \\ &= 1. \end{aligned}$$

Suppose $X \sim U[a, b]$. Then for every $x \in [a, b]$,

$$\mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \frac{x-a}{b-a}$$

Otherwise, if $x > b$, then $\mathbb{P}(X \leqslant x) = 1$, and if $x < a$, then $\mathbb{P}(X \leqslant x) = 0$.

Note that for $U[0, 1]$, we have $\mathbb{P}(X \leqslant x) = x$ for all $x \in [0, 1]$.

4.3 Exponential Distribution

Definition 4.6 (Exponential Distribution)

An **exponential distribution** with parameter $\lambda > 0$ is a continuous random variable X, denoted $\mathrm{Exp}(\lambda)$, with probability density function

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & x \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

> ***Proof of density validity.*** We have

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= \int_0^{\infty} \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \lambda \int_0^{\infty} \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \left[\lambda\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\right]_0^{\infty} \\ &= 1.\end{aligned}$$

Suppose $X \sim \mathrm{Exp}(\lambda)$. Then for every $x \geqslant 0$,

$$\mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = 1 - \mathrm{e}^{-\lambda x}$$
$$\mathbb{P}(X \geqslant x) = 1 - \mathbb{P}(X < x) = \mathrm{e}^{-\lambda x}.$$

Remark. Let $T \sim \mathrm{Exp}(\lambda)$, and $T_n = \lfloor nT \rfloor$ with $n \in \mathbb{N}$. Then, for every $k \in \mathbb{N}$,

$$\mathbb{P}(T_n \geqslant k) = \mathbb{P}(nT \geqslant k) = \mathbb{P}\left(T \geqslant \frac{k}{n}\right) = \left(\mathrm{e}^{-\frac{\lambda}{n}}\right)^k.$$

Note that T_n is a geometric random variable with parameter $p = 1 - \mathrm{e}^{-\frac{\lambda}{n}}$. As $n \to \infty$, $p \approx \frac{\lambda}{n}$.

Hence, $\frac{T_n}{n}$ converges in distribution to T as $n \to \infty$. So, we can think of an exponential distribution as a continuous analogue of a geometric distribution.

Proposition 4.7 (Memoryless Property of the Exponential Distribution)

Let $T \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$. Let $s, t \in \mathbb{R}_+$. Then

$$\begin{aligned}\mathbb{P}(T \geqslant s + t \mid T \geqslant s) &= \frac{\mathbb{P}(T \geqslant s + t, T \geqslant s)}{\mathbb{P}(T \geqslant s)} \\ &= \frac{\mathbb{P}(T \geqslant s + t)}{\mathbb{P}(T \geqslant s)} \\ &= \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}} \\ &= \mathrm{e}^{-\lambda t} \\ &= \mathbb{P}(T \geqslant t).\end{aligned}$$

It is significant that the exponential distribution is the only continuous distribution with the memoryless property.

Theorem 4.8

Let T be a positive random variable which is not identically zero or ∞. Then T has the exponential distribution iff T has the memoryless property, i.e. for all $s, t \in \mathbb{R}_+$,

$$\mathbb{P}(T \geqslant s + t \mid T \geqslant s) = \mathbb{P}(T \geqslant t).$$

Proof.

[⇒] This is shown in Memoryless Property of the Exponential Distribution 4.7.

[⇐] Suppose that T has the memoryless property. Let $t \in \mathbb{R}_+$ and

$$g(t) = \mathbb{P}(T \geqslant t).$$

Then $g(t+s) = g(t)g(s)$. For every $m \in \mathbb{N}$, $g(mt) = (g(t))^m$.

In particular, when $t = 1$, we have $g(m) = (g(1))^m$.

Let $g(1) = \mathbb{P}(T \geqslant 1) = \mathrm{e}^{-\lambda}$. We have

$$\lambda = -\log \mathbb{P}(T \geqslant 1),$$

and so $g(m) = \mathrm{e}^{-\lambda m}$.

By our definition, for any $m, n \in \mathbb{N}$,

$$\left(g\left(\frac{m}{n}\right)\right)^n = g(m) = \mathrm{e}^{-\lambda m},$$

so $g\left(\frac{m}{n}\right) = \mathrm{e}^{-\frac{\lambda m}{n}}$. Therefore, $g(r) = \mathrm{e}^{-\lambda r}$ for all $r \in \mathbb{Q}_+$. We shall extend this to $\mathbb{R}_+$.

Let $t > 0$, and $r, s \in \mathbb{Q}_+$ such that $s < t < r$ and $r - s \leqslant \varepsilon$. Then

$$\mathrm{e}^{-\lambda r} = g(r) \leqslant g(t) \leqslant g(s) = \mathrm{e}^{-\lambda s}.$$

Taking $\varepsilon \to 0$, we have $g(t) = \mathrm{e}^{-\lambda t}$.

Lecture 16 · 2026-02-27

4.4 Expectation and Variance of a Continuous Random Variable

Definition 4.9 (Expectation of a Continuous Random Variable)

Let X be a continuous random variable with density f with $X \geqslant 0$. The **expectation** of X is defined as

$$\mathbb{E}[X] = \int_0^\infty x f(x)\,\mathrm{d}x.$$

Let g be a non-negative function on $\mathbb{R}$. Then, we define

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

For a general random variable X, define $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$. Then, $X = X_+ - X_-$. If $\mathbb{E}[X_+] < \infty$ or $\mathbb{E}[X_-] < \infty$, then we define

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

Remark. As in the discrete case, we have

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

Lemma 4.10

If X is a continuous random variable with $X \geqslant 0$, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geqslant x)\, dx.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f(x)\, dx \\
&= \int_0^\infty \left(\int_0^x dt \right) f(x)\, dx \\
&= \int_0^\infty \left(\int_t^\infty f(x)\, dx \right) dt \\
&= \int_0^\infty (1 - F(t))\, dt \\
&= \int_0^\infty \mathbb{P}(X \geqslant t)\, dt.
\end{aligned}$$

Recall that in the discrete case, we have

$$X = \sum_{n=1}^{\infty} \mathbb{1}(X \geqslant n).$$

Lemma 4.11

If X is a continuous random variable with $X \geqslant 0$, then for any outcome ω,

$$X(\omega) = \int_0^\infty \mathbb{1}(X(\omega) \geqslant x)\, dx,$$

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{1}(X \geqslant x)\, dx\right] = \int_0^\infty \mathbb{P}(X \geqslant x)\, dx.$$

Definition 4.12 (Variance of a Continuous Random Variable)

Let X be a continuous random variable with $\mathbb{E}\left[X^2\right] < \infty$. The **variance** of X is defined as

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

Example 4.13

1. Consider $X \sim U[a,b]$. Then the density f of X is given by

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

Hence

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \frac{a+b}{2}.$$

2. Consider $X \sim \text{Exp}(\lambda)$. Then for $\lambda > 0$, $f(x) = \lambda \mathrm{e}^{-\lambda x}$ and $x > 0$. Hence

$$\begin{aligned} \mathbb{E}[X] &= \int_0^{\infty} x\lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \int_0^{\infty} \mathbb{P}(X \geqslant x)\,\mathrm{d}x \\ &= \int_0^{\infty} \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \frac{1}{\lambda}. \end{aligned}$$

4.5 Normal Distribution

4.5.1 Introduction

Definition 4.14 (Normal Distribution)

An **normal distribution** with parameters $-\infty < \mu < \infty$ and $\sigma \in \mathbb{R}_+$ is a continuous random variable X, denoted $N\left(\mu, \sigma^2\right)$, with probability density function

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

Proof of density validity. We have

$$\begin{aligned} I &= \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \qquad \text{with } u = \frac{x-\mu}{\sigma} \\ &= 2\int_{0}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \end{aligned}$$

Consider I^2. We have

$$I^2 = \frac{2}{\pi}\int_0^\infty \int_0^\infty \exp\left(-\frac{u^2}{2}\right)\exp\left(-\frac{v^2}{2}\right) du\, dv$$

Changing to polar coordintates with $u = r\cos(\theta)$ and $v = r\sin(\theta)$, we have

$$\begin{aligned} I^2 &= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} r e^{-\frac{r^2}{2}}\, d\theta\, dr \\ &= \int_0^\infty r e^{-\frac{r^2}{2}}\, dr \\ &= \left[-e^{-\frac{r^2}{2}}\right]_0^\infty \\ &= 1. \end{aligned}$$

Since $I > 0$, we have $I = 1$. Hence, f is a valid probability density function.

Proposition 4.15

Let $X \sim N\left(\mu, \sigma^2\right)$ with $-\infty < \mu < \infty$ and $\sigma \in \mathbb{R}_+$. Then

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

Proof. We have

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^\infty x f(x)\, dx \\ &= \int_{-\infty}^\infty (x-\mu)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx + \underbrace{\int_{-\infty}^\infty \mu\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx}_{\mu\int_{-\infty}^\infty f(x)\,dx = \mu} \\ &= \int_{-\infty}^\infty \frac{u}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{u^2}{2\sigma^2}\right)dx + \mu && \text{with } u = x-\mu \\ &= \mu. && \text{the integrand is odd} \end{aligned}$$

Hence, $\mathbb{E}[X] = \mu$. Moreover,

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= \int_{-\infty}^\infty (x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \int_{-\infty}^\infty \sigma^2 u^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{u^2}{2}\right)du && \text{with } u = \frac{x-\mu}{\sigma} \\ &= \sigma^2. \end{aligned}$$

Hence $\mathrm{Var}(X) = \sigma^2$.

4.5.2 Linear Transformations of Normal Distributions

Theorem 4.16

Let X have density f, and let g be a function which is strictly monotone and g^{-1} is differentiable. Then $g(X)$ has density

$$f\left(g^{-1}(x)\right) \cdot \left|\left(g^{-1}\right)'(x)\right|.$$

Proof.

- Suppose that g is strictly increasing. Then

$$\begin{aligned}\mathbb{P}(g(X) \leqslant x) &= \mathbb{P}\left(X \leqslant g^{-1}(x)\right) \\ &= F\left(g^{-1}(x)\right) \\ \frac{\mathrm{d}}{\mathrm{d}x}(\mathbb{P}(g(X) \leqslant x)) &= f\left(g^{-1}(x)\right) \cdot \left(g^{-1}\right)'(x)\end{aligned}$$

- Suppose that g is strictly decreasing. Then

$$\begin{aligned}\mathbb{P}(g(X) \leqslant x) &= \mathbb{P}\left(X \geqslant g^{-1}(x)\right) \\ &= 1 - \mathbb{P}\left(X < g^{-1}(x)\right) \\ \frac{\mathrm{d}}{\mathrm{d}x}(\mathbb{P}(g(X) \leqslant x)) &= -f\left(g^{-1}(x)\right) \cdot \left(g^{-1}\right)'(x)\end{aligned}$$

Hence the result follows in either case.

Proposition 4.17 (Linear Transformations of Normal Distributions)

Let $X \sim N\left(\mu, \sigma^2\right)$ with $-\infty < \mu < \infty$ and $\sigma \in \mathbb{R}_+$. Let $a, b \in \mathbb{R}$ with $a \neq 0$. Then

$$aX + b \sim N\left(a\mu + b, (a\sigma)^2\right).$$

Proof. Define $g(x) = ax + b$ and $Y = g(X)$. We have $g^{-1}(x) = \frac{x-b}{a}$ and $\left(g^{-1}\right)'(x) = \frac{1}{a}$. Hence,

$$\begin{aligned}f_Y(y) &= f_X\left(g^{-1}(y)\right) \cdot \left|\left(g^{-1}\right)'(y)\right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}\right) \cdot \frac{1}{|a|} \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}} \exp\left(-\frac{(y - (a\mu + b))^2}{2a^2\sigma^2}\right).\end{aligned}$$

So $Y \sim N\left(a\mu + b, (a\sigma)^2\right)$.

Remark. If $X \sim N\left(\mu, \sigma^2\right)$, then

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

Example 4.18

Let $X \sim N\left(\mu, \sigma^2\right)$, and consider

$$\mathbb{P}(-2\sigma < X-\mu < 2\sigma) = \mathbb{P}\left(-2 < \frac{X-\mu}{\sigma} < 2\right).$$

Let

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = \mathbb{P}(Z \leqslant x).$$

Note that

$$\varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$(\Phi(x) + \Phi(-x))' = 0.$$

Since $\Phi(0) = \frac{1}{2}$, we have $\Phi(x) + \Phi(-x) = 1$. Hence,

$$\mathbb{P}(X \leqslant x) + \mathbb{P}(X \leqslant -x) = 1.$$

We have existing tables of Φ values, so we can compute $\mathbb{P}(X \leqslant x)$ for any x, in particular,

$$\mathbb{P}(-2\sigma < X-\mu < 2\sigma) = \mathbb{P}(-2 < Z < 2) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 > 0.95.$$

Lecture 17 · 2026-03-02

Definition 4.19 (Median)

Let X be a continuous random variable. The median of X, denoted by m, is the value such that

$$\mathbb{P}(X \geqslant m) = \mathbb{P}(X \leqslant m) = \frac{1}{2}.$$

Alternatively,

$$\int_{\infty}^{m} f(x)\,dx = \int_{m}^{\infty} f(x)\,dx = \frac{1}{2}$$

where f is the density of X.

Example 4.20

Suppose that $X \sim N\left(\mu, \sigma^2\right)$. Then

$$\mathbb{P}(X \leqslant \mu) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leqslant 0\right) = \mathbb{P}(N(0,1) \leqslant 0) = \frac{1}{2}.$$

Hence the median of X is μ.

4.6 Multivariate Density Functions

4.6.1 Introduction

Definition 4.21 (Multivariate Density Function)

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ be a random vector. We say that $\boldsymbol{X}$ has a **multivariate density function** if there exists a non-negative function $f : \mathbb{R}^n \to \mathbb{R}$ such that for all $x_1, ..., x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(y_1, ..., y_n)\, dy_n ... dy_1.$$

The **probability distribution function** F of $\boldsymbol{X}$ is defined as

$$F(x_1, ..., x_n) = \mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n).$$

Therefore,

$$f(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 ... \partial x_n} F(x_1, ..., x_n).$$

More generally, for $B \subseteq \mathbb{R}^n$,

$$\mathbb{P}((X_1, ..., X_n)^\top \in B) = \int_B f(y_1, ..., y_n)\, dy_1 ... dy_n.$$

Definition 4.22 (Independence of Continuous Random Variables)

Let $X_1, ..., X_n$ be continuous random variables. They are **independent** if for all $x_1, ..., x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n) = \mathbb{P}(X_1 \leqslant x_1) ... \mathbb{P}(X_n \leqslant x_n).$$

Theorem 4.23

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ be a random vector with density f.

1. Suppose $X_1, ..., X_n$ are independent with densities $f_1, ..., f_n$, then for all $x_1, ..., x_n \in \mathbb{R}$,
$$f(x_1, ..., x_n) = f_1(x_1) ... f_n(x_n).$$
2. Conversely, suppose f factorises as $f(x_1, ..., x_n) = f_1(x_1) ... f_n(x_n)$ for some non-negative functions $f_1, ..., f_n$ on $\mathbb{R}$. Then $X_1, ..., X_n$ are independent with densities proportional to $f_1, ..., f_n$.

Proof.

1. We have

$$\begin{aligned}\mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n) &= \mathbb{P}(X_1 \leqslant x_1)...\mathbb{P}(X_n \leqslant x_n) \\ &= \left(\int_{-\infty}^{x_1} f_1(y_1)\,dy_1\right)...\left(\int_{-\infty}^{x_n} f_n(y_n)\,dy_n\right) \\ &= \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_1(y_1)...f_n(y_n)\,dy_n...\,dy_1.\end{aligned}$$

So $f(x_1, ..., x_n) = f_1(x_1)...f_n(x_n)$.

2. We have

$$\begin{aligned}\mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n) &= \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(y_1, ..., y_n)\,dy_n...\,dy_1 \\ &= \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_1(y_1)...f_n(y_n)\,dy_n...\,dy_1 \\ &= \left(\int_{-\infty}^{x_1} f_1(y_1)\,dy_1\right)...\left(\int_{-\infty}^{x_n} f_n(y_n)\,dy_n\right).\end{aligned}$$

Note that

$$\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(y_1)...f_n(y_n)\,dy_n...\,dy_1 = 1.$$

Hence

$$\mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n) = \frac{\int_{-\infty}^{x_1} f_1(y_1)\,dy_1}{\int_{-\infty}^{\infty} f_1(y_1)\,dy_1}...\frac{\int_{-\infty}^{x_n} f_n(y_n)\,dy_n}{\int_{-\infty}^{\infty} f_n(y_n)\,dy_n}.$$

So, $X_1, ..., X_n$ are independent with densities proportional to $f_1, ..., f_n$.

4.6.2 Marginal Density Functions

Definition 4.24 (Marginal Density Function)

Let $\boldsymbol{X} = (X_1, ..., X_n)^\top$ be a random vector with density f. The **marginal density function** of X_1 is defined as

$$f_{X_1}(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x, x_2, ..., x_n)\,dx_n...\,dx_2.$$

Proof. Suppose $\boldsymbol{X} = (X_1, ..., X_n)^\top$ has density f. Then consider

$$\begin{aligned}\mathbb{P}(X_1 \leqslant x) &= \mathbb{P}(X_1 \leqslant x, X_2 \in \mathbb{R}, ..., X_n \in \mathbb{R}) \\ &= \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, ..., x_n)\,dx_n...\,dx_2\right) dx_1.\end{aligned}$$

So the density of X_1 is

$$f_{X_1}(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x, x_2, ..., x_n)\,dx_n...\,dx_2.$$

4.6.3 Sum of Independent Random Variables

Suppose X and Y are independent random variables with densities f_X and f_Y respectively. We want to find the density of $X+Y$.

Recall that in the discrete case, we used the discrete convolution formula

$$\mathbb{P}(X+Y=z)=\sum_{x\in\mathbb{R}}\mathbb{P}(X=x,Y=z-x).$$

For the continuous case, we have the following analogue of the discrete convolution formula.

$$f_{X+Y}(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\,\mathrm{d}x.$$

Proof. We have

$$\begin{aligned}\mathbb{P}(X+Y\leqslant z)&=\mathbb{P}\Big((X,Y)\in\big\{(x,y)\in\mathbb{R}^2:x+y\leqslant z\big\}\Big)\\&=\int_{\{x+y\leqslant z\}}f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\\&=\int_{\{x+y\leqslant z\}}f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y\\&=\int_{-\infty}^{\infty}f_X(x)\int_{-\infty}^{z-x}f_Y(y)\,\mathrm{d}y\,\mathrm{d}x\\&=\int_{-\infty}^{\infty}f_X(x)\int_{-\infty}^{z}f_Y(y-x)\,\mathrm{d}y\,\mathrm{d}x\\&=\int_{-\infty}^{z}\int_{-\infty}^{\infty}f_X(x)f_Y(y-x)\,\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

So the density of $X+Y$ is given by the convolution formula

$$f_{X+Y}(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\,\mathrm{d}x.$$

4.6.4 Conditional Density Functions

Definition 4.25 (Conditional Density Function)

Let X and Y be two random variables with joint density $f_{X,Y}$ and marginal densities f_X and f_Y respectively.

The **conditional density function** of X given $Y=y$ is defined as

$$f_{X|Y}(x\mid y)=\frac{f(X,Y)(x,y)}{f_Y(y)}\quad\text{for } f_Y(y)>0.$$

Proposition 4.26 (Law of Total Probability for Continuous Random Variables)

Let X and Y be two random variables with joint density $f_{X,Y}$ and marginal densities f_X and f_Y respectively. Then, for every $x \in \mathbb{R}$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}y.$$

Similarly, we can define the **conditional expectation** of X given Y.

Definition 4.27 (Conditional Expectation of a Continuous Random Variable)

The **conditional expectation** of X given Y is defined as $\mathbb{E}[X \mid Y] = g(Y)$, where

$$g(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

4.6.5 Transformation of Random Variables

Theorem 4.28

Let $\boldsymbol{X}$ be a random variable with values in $D \subseteq \mathbb{R}^n$ and density f. Let $g : D \to g(D)$ be a bijection with a continuous derivative on D, and

$$\det g'(\boldsymbol{x}) \neq 0 \quad \text{for all } \boldsymbol{x} \in D.$$

Then, the random variable $Y = g(\boldsymbol{X})$ has density

$$f_Y(\boldsymbol{y}) = f_X\left(g^{-1}(\boldsymbol{y})\right) \cdot |J|$$

where $J = \det\left(\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1}^{n}\right)$ is the Jacobian determinant of g^{-1}.

Lecture 18 · 2026-03-04

Example 4.29

Let $X, Y \sim N(0,1)$ be independent. Then, consider (R, Θ) in polar coordinates with $R = \sqrt{X^2 + Y^2}$ and $\Theta = \arctan\left(\frac{Y}{X}\right)$. We have

$$\begin{aligned} X &= R\cos\Theta \\ Y &= R\sin\Theta. \end{aligned}$$

We want to find the density of (R, Θ). We have

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta) \cdot |\det \mathbf{J}|$$

where

$$\mathbf{J} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

and $\det \mathbf{J} = r$. Hence,

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= f_X(r\cos\theta)\cdot f_Y(r\sin\theta)\cdot r \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{r^2\cos^2\theta}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{r^2\sin^2\theta}{2}\right)\cdot r \\ &= \frac{1}{2\pi}e^{-\frac{r^2}{2}}\cdot r \end{aligned}$$

where $r \geqslant 0$ and $\theta \in [0, 2\pi]$.

Hence, $\Theta \sim U[0,2\pi]$ and R has density $f_R(r) = re^{-\frac{r^2}{2}}$ for $r \geqslant 0$, and they are independent, by Theorem 4.23 (2).

4.7 Order Statistics of a Random Sample

Definition 4.30 (Order Statistics)

Let $X_1, ..., X_n$ be i.i.d. random variables with probability distribution function F and density f. Order them from the smallest to the largest as

$$X_{(1)} \leqslant X_{(2)} \leqslant ... \leqslant X_{(n)}.$$

Lete $Y_i = X_{(i)}$. Then, $Y_1, ..., Y_n$ are called the **order statistics** of the random sample $X_1, ..., X_n$.

We aim to find the density of $(Y_1, ..., Y_n)$. For the minimum Y_1, we have

$$\begin{aligned} \mathbb{P}(Y_1 \leqslant x) &= 1 - \mathbb{P}(Y_1 > x) \\ &= 1 - \mathbb{P}(X_1 > x, ..., X_n > x) \\ &= 1 - (1 - F(x))^n \\ f_{Y_1}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(Y_1 \leqslant x) = n(1-F(x))^{n-1}f(x). \end{aligned}$$

For the maximum Y_n, we have

$$\begin{aligned} \mathbb{P}(Y_n \leqslant x) &= \mathbb{P}(X_1 \leqslant x)^n = (F(x))^n \\ f_{Y_n}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(Y_n \leqslant x) = n(F(x))^{n-1}f(x). \end{aligned}$$

In order to find $f_{Y_1,...,Y_n}(x_1, ..., x_n)$ with $x_1 < x_2 < ... < x_n$, we have

$$\begin{aligned} \mathbb{P}(Y_1 \leqslant x_1, ..., Y_n \leqslant x_n) &= n!\mathbb{P}(X_1 \leqslant x_1, ..., X_n \leqslant x_n, X_1 \leqslant X_2 \leqslant ... \leqslant X_n) \\ &= n!\int ... \int f^{\top}_{X_1,...,X_n}(u_1, ..., u_n)\mathbb{1}(u_1 \leqslant x_1, ..., u_n \leqslant x_n, u_1 \leqslant u_2 \leqslant ... \leqslant u_n)\,\mathrm{d}u_n ... \mathrm{d}u_1 \\ &= n!\int_{-\infty}^{x_1}\int_{u_1}^{x_2} ... \int_{u_{n-1}}^{x_n} f(u_1)f(u_2)...f(u_n)\,\mathrm{d}u_n ... \mathrm{d}u_1. \end{aligned}$$

Hence,

$$f_{Y_1,\dots,Y_n}(x_1,\dots,x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n}\mathbb{P}(Y_1 \leqslant x_1,\dots,Y_n \leqslant x_n)$$
$$= \begin{cases} n! f(x_1) f(x_2) \dots f(x_n) & \text{for } x_1 < x_2 < \dots < x_n \\ 0 & \text{otherwise} \end{cases}$$

4.7.1 Order Statistics of Exponential Distributions

Let $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$, where $X \perp Y$ and $\lambda, \mu > 0$. Let $Z = \min(X, Y)$. Then

$$\mathbb{P}(Z \leqslant z) = 1 - \mathbb{P}(Z > z) = 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z)$$
$$= 1 - e^{-\lambda z} e^{-\mu z} = 1 - e^{-(\lambda+\mu)z} \quad \text{for } z > 0.$$

Hence, $Z \sim \text{Exp}(\lambda + \mu)$.

If (X_i) are independent random variables with $X_i \sim \text{Exp}(\lambda_i)$, then $\min(X_1,\dots,X_n)^\top \sim \text{Exp}(\sum_{i=1}^n \lambda_i)$.

Now, consider $X_1, X_2, \dots, X_n$ be i.i.d. random variables with $X_i \sim \text{Exp}(\lambda)$. Let $Y_i = X_{(i)}$ be the order statistics of $X_1, \dots, X_n$.

Let $Z_1 = Y_1, Z_2 = Y_2 - Y_1, \dots, Z_n = Y_n - Y_{n-1}$. Note that we have found the distribution of Z_1 above.

Consider the joint density of $(Z_1, \dots, Z_n)$. We have

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \mathbf{A}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{where } \mathbf{A} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Hence, with the transformation of $z = \mathbf{A}y$ and $y_j = \sum_{i=1}^j z_i$, we have

$$f_{Z_1,\dots,Z_n}(z_1,\dots,z_n) = f_{Y_1,\dots,Y_n}(y_1,\dots,y_n)|\det \mathbf{J}| \quad \text{where } \mathbf{J} = \mathbf{A}^{-1}$$
$$= n! f(y_1) \dots f(y_n)$$
$$= n! \lambda e^{-\lambda y_1} \dots \lambda e^{-\lambda y_n}$$
$$= \prod_{i=1}^n \left((n-i+1)\lambda e^{-\lambda(n-i+1)z_i} \right).$$

Hence, $Z_i \sim \text{Exp}((n-i+1)\lambda)$ and they are independent, by Theorem 4.23 (2).

4.8 Moment Generating Functions

Definition 4.31 (Moment Generating Function)

Let X be a random variable with density f. The **moment generating function** (MGF) of X is

$$m(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx$$

whenever the integral is finite. Note that $m(0) = 1$.

Theorem 4.32

The MGF uniquely determines the distribution of a random variable, provided it is defined for an open interval of values of θ.

Theorem 4.33

Suppose that the MGF is defined for an interval of values of θ, then

$$m^{(r)}(0) = \left.\frac{\mathrm{d}^r m(\theta)}{\mathrm{d}\theta^r}\right|_{\theta=0} = \mathbb{E}[X^r].$$

4.8.1 Gamma Distribution

Example 4.34 (Gamma Distribution)

For $n \in \mathbb{N}$ and $\lambda > 0$, the **gamma distribution** with parameters n and λ is a continuous random variable X with density

$$f(x) = \frac{\mathrm{e}^{-\lambda x} \cdot \lambda^n \cdot x^{n-1}}{(n-1)!} \quad \text{for } x > 0.$$

Proof of density validity. We have

$$\begin{aligned} I_n &= \int_0^\infty f(x)\,\mathrm{d}x = \int_0^\infty \lambda \mathrm{e}^{-\lambda x} \frac{\lambda^{n-1} x^{n-1}}{(n-1)!}\,\mathrm{d}x \\ &= \int_0^\infty \lambda \mathrm{e}^{-\lambda x} \frac{\lambda^{n-1} x^{n-1}}{(n-1)!}\,\mathrm{d}x \\ &= \int_0^\infty \lambda \mathrm{e}^{-\lambda x} \frac{\lambda^{n-2} x^{n-2}}{(n-2)!}\,\mathrm{d}x \\ &= I_{n-1} = \ldots = I_1 = \int_0^\infty \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = 1. \end{aligned}$$

We denote $X \sim \Gamma(n, \lambda)$. Then,

$$\begin{aligned} m(\theta) = \mathbb{E}\left[\mathrm{e}^{\theta x}\right] &= \int_0^\infty \mathrm{e}^{\theta x} \cdot \mathrm{e}^{-\lambda x} \cdot \frac{\lambda^n \cdot x^{n-1}}{(n-1)!}\,\mathrm{d}x \\ &= \underbrace{\int_0^\infty \mathrm{e}^{-(\lambda-\theta)} \frac{(\lambda-\theta)^n \cdot x^{n-1}}{(n-1)!}\,\mathrm{d}x}_{=1} \cdot \frac{\lambda^n}{(\lambda-\theta)^n}. \end{aligned}$$

If $\theta < \lambda$, then $m(\theta) = \left(\frac{\lambda}{\lambda-\theta}\right)^n$.

Lecture 19 · 2026-03-06

Proposition 4.35

If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$m(\theta) = \mathbb{E}\left[e^{\theta(X_1+X_2+\ldots+X_n)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\theta X_i}\right].$$

Suppose $X \sim \Gamma(n, \lambda)$ and $Y \sim \Gamma(m\lambda)$, where $m, n \in \mathbb{N}$, $\lambda > 0$ and $X \perp Y$. Consider the density of $X + Y$.

We aim to show this by Theorem 4.32. Consider, for $\theta < \lambda$,

$$\begin{aligned}\mathbb{E}\left[e^{\theta(X+Y)}\right] &= \mathbb{E}\left[e^{\theta X}\right]\mathbb{E}\left[e^{\theta Y}\right] \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \left(\frac{\lambda}{\lambda-\theta}\right)^m \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^{n+m}.\end{aligned}$$

Hence $X + Y \sim \Gamma(n + m, \lambda)$

Suppose $X_1, \ldots, X_n$ are i.i.d. with $X_i \sim \text{Exp}(\lambda)$. Then, $X_1 + \ldots + X_n \sim \Gamma(n, \lambda)$.

Remark. One could also define $\Gamma(\alpha, \lambda)$ with $\alpha, \lambda > 0$ by replacing $(n-1)!$ in the density by

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

We say $X \sim \Gamma(\alpha, \lambda)$ if $f(x) = \frac{e^{-\lambda x} \cdot \lambda^\alpha \cdot x^{\alpha-1}}{\Gamma(\alpha)}$ for $x > 0$.

Cauchy Distribution (Non-Examinable). The Cauchy distribution is defined as the distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } x \in \mathbb{R}.$$

The MGF is

$$m(\theta) = \int_{-\infty}^{\infty} \frac{e^{\theta x}}{\pi(1+x^2)}\, dx = \begin{cases} 1 & \text{for } \theta = 0 \\ \infty & \text{otherwise} \end{cases}$$

Hence, $X, 2X, 3X, \ldots$ all have the same MGF. However, it is not the case that $X, 2X, 3X, \ldots$ all have the same distribution. So the assumption that the MGF must be finite for an open interval of values of θ is necessary in Theorem 4.32.

4.8.2 MGF of the Normal Distribution

Recall that if $X \sim N\left(\mu, \sigma^2\right)$, then

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

Hence, the MGF of X is

$$m(\theta) = \mathbb{E}\left[e^{\theta x}\right] = \int_{-\infty}^{\infty} e^{\theta x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

Note that, the exponent is

$$\begin{aligned}
\text{exponent} &= \theta x - \frac{(x-\mu)^2}{2\sigma^2} \\
&= \theta x - \frac{x^2}{2\sigma^2} + \frac{2x\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} \\
&= -\frac{x^2}{2\sigma^2} + 2\frac{x}{2\sigma^2}\left(\mu + \theta\sigma^2\right) - \frac{\mu^2}{2\sigma^2} \\
&= -\frac{x^2}{2\sigma^2} + 2\frac{x}{2\sigma^2}\left(\mu + \theta\sigma^2\right) - \frac{\left(\mu+\theta\sigma^2\right)^2}{2\sigma^2} + \frac{\left(\mu+\theta\sigma^2\right)^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} \\
&= -\frac{1}{2\sigma^2}\left(x - \left(\mu+\theta\sigma^2\right)\right)^2 + \cancel{\frac{\mu^2}{2\sigma^2}} + \frac{2\mu\theta\sigma^2}{2\sigma^2} + \frac{\theta^2\sigma^2}{2} - \cancel{\frac{\mu^2}{2\sigma^2}}
\end{aligned}$$

Hence,

$$\begin{aligned}
m(\theta) &= \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\left(\mu+\theta\sigma^2\right)\right)^2}{2\sigma^2}\right) dx}_{\text{integral of density of } N(\mu+\theta\sigma^2,\sigma^2)} \cdot \exp\left(\mu\theta + \frac{\theta^2\sigma^2}{2}\right) \\
&= \exp\left(\mu\theta + \frac{\theta^2\sigma^2}{2}\right).
\end{aligned}$$

Let $X \sim N\left(\mu, \sigma^2\right)$ and $Y \sim N\left(\nu, \tau^2\right)$, and $X \perp Y$. Then

$$\begin{aligned}
m(\theta) = \mathbb{E}\left[e^{\theta(X+Y)}\right] &= \exp\left(\mu\theta + \frac{\theta^2\sigma^2}{2}\right)\exp\left(\nu\theta + \frac{\theta^2\tau^2}{2}\right) \\
&= \exp\left((\mu+\nu)\theta + \frac{\theta^2\left(\sigma^2+\tau^2\right)}{2}\right).
\end{aligned}$$

Hence $X + Y \sim N\left(\mu+\nu, \sigma^2+\tau^2\right)$.

4.8.3 Multivariate Moment Generating Functions

Definition 4.36 (Multivariate Moment Generating Function)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^n$ be a random variable. The MGF of X is defined to be

$$m(\theta) = \mathbb{E}\left[e^{\theta^\top \cdot X}\right] = \mathbb{E}\left[e^{\sum_{i=1}^n \theta_i X_i}\right],$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$.

Theorem 4.37

For a multivariate random variable, if the MGF is finite for an open set of values of $\boldsymbol{\theta}$, then it uniquely determines the distribution of the random variable.

In this case,

$$\left.\frac{\partial^r m}{\partial \theta_i^r}\right|_{\theta=0} = \mathbb{E}[X_i^r] \quad \text{and} \quad \left.\frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s}\right|_{\theta=0} = \mathbb{E}\left[X_i^r X_j^s\right].$$

Proposition 4.38

Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ be a random variable in $\mathbb{R}^n$. Then

$$m(\theta) = \mathbb{E}\left[e^{\theta^\top X}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\theta_i X_i}\right]$$

iff $X_1, \dots, X_n$ are independent.

Proof.

[⇐] This is a direct consequence of the definition of MGF.

[⇒] If $X_1, \dots, X_n$ are independent, then $m(\theta)$ factorises. If $m(\theta)$ factorises, then by Theorem 4.32, $X_1, \dots, X_n$ are independent.

4.9 Multidimensional Gaussian Random Variables

4.9.1 Introduction

Definition 4.39 (Gaussian Random Variable)

A random variable X with values in $\mathbb{R}$ is called **Gaussian** (or **normal**) in $\mathbb{R}$ if it can be written as

$$X \sim \mu + \sigma Z$$

where $Z \sim N(0,1)$, $\mu \in \mathbb{R}$, $\sigma \geqslant 0$.

If $\sigma > 0$, then the density of X is, for $x \in \mathbb{R}$,

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

Definition 4.40 (Gaussian Vector)

Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top \in \mathbb{R}^n$ be a random variable. We say that $\boldsymbol{X}$ is a **Gaussian vector** (or **Gaussian in $\mathbb{R}^n$**) if for all $\boldsymbol{u} \in \mathbb{R}^n$,

$$\boldsymbol{u}^\top \boldsymbol{X} = \sum_{i=1}^{n} u_i X_i$$

is a Gaussian random variable in $\mathbb{R}$.

Proposition 4.41

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be a Gaussian vector. Let $\mathbf{A}$ be an $m \times n$ matrix and $b \in \mathbb{R}^m$. Then $\mathbf{A}\boldsymbol{X} + \boldsymbol{b}$ is also a Gaussian vector.

Proof. Let $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$. We need to show that $\boldsymbol{u}^\top(\mathbf{A}\boldsymbol{X} + \boldsymbol{b})$ is a Gaussian random variable in $\mathbb{R}$. Note that

$$\boldsymbol{u}^\top(\mathbf{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{u}^\top\mathbf{A}\boldsymbol{X} + \boldsymbol{u}^\top\boldsymbol{b}$$

Letting $\boldsymbol{v} = \mathbf{A}^\top\boldsymbol{u}$, we have

$$\begin{aligned}\boldsymbol{u}^\top(\mathbf{A}\boldsymbol{X} + \boldsymbol{b}) &= \boldsymbol{v}^\top\boldsymbol{X} + \boldsymbol{u}^\top\boldsymbol{b} \\ &= \boldsymbol{v}^\top\boldsymbol{X} + \sum_{i=1}^n u_i b_i.\end{aligned}$$

Since $\boldsymbol{X}$ is a Gaussian vector, $\boldsymbol{v}^\top\boldsymbol{X}$ is a Gaussian random variable in $\mathbb{R}$. Note that $\sum_{i=1}^n u_i b_i$ is a constant. Hence, $\boldsymbol{u}^\top(\mathbf{A}\boldsymbol{X} + \boldsymbol{b})$ is also a Gaussian random variable in $\mathbb{R}$.

Definition 4.42

Define

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = \begin{pmatrix}\mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n]\end{pmatrix}$$

$$\mathbf{V} = \mathrm{Var}(\boldsymbol{X}) = \mathbb{E}\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\big].$$

Note that

$$((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top)_{ij} = (X_i - \mu_i)\big(X_j - \mu_j\big).$$

$$\mathrm{Var}(\boldsymbol{X})_{ij} = \mathrm{Cov}\big(X_i, X_j\big).$$

Hence, $\mathrm{Var}(\boldsymbol{X})$ is a symmetric matrix.

Lecture 20 · 2026-03-09

Consider the random variable $\boldsymbol{u}^\top\boldsymbol{X}$ for some $\boldsymbol{u} \in \mathbb{R}^n$. We have

$$\mathbb{E}\big[\boldsymbol{u}^\top\boldsymbol{X}\big] = \mathbb{E}\left[\sum_{i=1}^n u_i X_i\right] = \sum_{i=1}^n u_i \mathbb{E}[X_i] = \boldsymbol{u}^\top\boldsymbol{\mu}$$

$$\begin{aligned}\mathrm{Var}(\boldsymbol{u}^\top\boldsymbol{X}) &= \mathrm{Var}\left(\sum_{i=1}^n u_i X_i\right) \\ &= \sum_{i,j=1}^n u_i u_j \,\mathrm{Cov}\big(X_i, X_j\big) \\ &= \boldsymbol{u}^\top\mathbf{V}\boldsymbol{u}.\end{aligned}$$

Proposition 4.43

$\mathbf{V}$ is a non-negative definite matrix, *i.e.* for any $\boldsymbol{u} \in \mathbb{R}^n$,

$$\boldsymbol{u}^\top \mathbf{V} \boldsymbol{u} \geqslant 0.$$

Proof. Note that $\operatorname{Var}(\boldsymbol{u}^\top \boldsymbol{X}) = \boldsymbol{u}^\top \mathbf{V} \boldsymbol{u}$. Since $\operatorname{Var}(\boldsymbol{u}^\top \boldsymbol{X}) \geqslant 0$, we have $\boldsymbol{u}^\top \mathbf{V} \boldsymbol{u} \geqslant 0$.

Consider the MGF of $\boldsymbol{X}$. We have

$$m(\lambda) = \mathbb{E}\left[e^{\lambda^\top \boldsymbol{X}}\right] \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^n$$

Note that $\boldsymbol{\lambda}^\top \boldsymbol{X}$ is $N(\boldsymbol{\lambda}^\top \boldsymbol{\mu}, \boldsymbol{\lambda}^\top \mathbf{V} \boldsymbol{\lambda})$. So

$$m(\lambda) = \exp\left(\boldsymbol{\lambda}^\top \boldsymbol{\mu} + \frac{1}{2} \boldsymbol{\lambda}^\top \mathbf{V} \boldsymbol{\lambda}\right).$$

[Recall that if $Z \sim N\left(\mu, \sigma^2\right)$, $\mathbb{E}\left[e^{\theta Z}\right] = \exp\left(\theta \mu + \frac{1}{2}\theta^2 \sigma^2\right)$.]

We have seen that the MGF uniquely characterises the distribution if defined for an open set of values. Hence, to characterise a Gaussian vector, we only need the mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$.

We have determined the MGF of a Gaussian vector purely from the definition of a Gaussian vector, and we will consider its density function later.

4.9.2 Construction of a Gaussian Random Vector

Lemma 4.44

Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. with $Z_i \sim N(0,1)$. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$. Then, $\boldsymbol{Z}$ is a Gaussian vector.

Proof. We need to show that $\forall \boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{u}^\top \boldsymbol{Z}$ is normal in $\mathbb{R}$.

The MGF of $\boldsymbol{u}^\top \boldsymbol{Z}$ is

$$m(\lambda) = \mathbb{E}\left[e^{\lambda \boldsymbol{u}^\top \boldsymbol{Z}}\right] = \mathbb{E}\left[e^{\sum_{i=1}^n \lambda u_i Z_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda u_i Z_i}\right] = \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^n u_i^2\right) = \exp\left(\frac{\lambda^2}{2} |\boldsymbol{u}|^2\right).$$

So $\boldsymbol{u}^\top \boldsymbol{Z} \sim N\left(0, |\boldsymbol{u}|^2\right)$, and hence $\boldsymbol{Z}$ is a Gaussian vector.

Remark. We have

$$\mathbb{E}[\boldsymbol{Z}] = \mathbf{0} \quad \text{and} \quad \operatorname{Var}(\boldsymbol{Z}) = \mathbf{I}_n.$$

We write that $\boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$

Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $\mathbf{V}$ be a non-negative definite matrix. We want to construct a Gaussian vector with mean $\boldsymbol{\mu}$ and (co)variance matrix $\mathbf{V}$, using the standard Gaussian vector $\boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$.

[Note that in the $n = 1$ case, we can construct $X \sim N\left(\mu, \sigma^2\right)$ by letting $X = \mu + \sigma Z$.]

Note that we will need some form of "square root" of $\mathbf{V}$.

Definition 4.45

Let $\mathbf{V}$ be a non-negative definite matrix. Consider

$$\mathbf{V} = \mathbf{U}^\top \mathbf{D} \mathbf{U} \quad \text{where } \mathbf{U}^\top = \mathbf{U}^{-1}$$

and $\mathbf{D}$ is a diagonal matrix with

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{pmatrix} \quad \text{where } \lambda_1, \dots, \lambda_n \geqslant 0.$$

Then, the **square root** of $\mathbf{V}$ is defined as

$$\boldsymbol{\sigma} = \mathbf{U}^\top \sqrt{\mathbf{D}} \mathbf{U} \quad \text{where } \sqrt{\mathbf{D}} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \sqrt{\lambda_n} \end{pmatrix}.$$

Note that $\boldsymbol{\sigma}^2 = \mathbf{V}$.

Lemma 4.46

Let $\boldsymbol{\mu} \in \mathbb{R}^n$, $\mathbf{V}$ be a non-negative definite matrix. Let $Z_1, Z_2, \dots, Z_n$ be i.i.d. with $Z_i \sim N(0,1)$, and let $\boldsymbol{Z} = (Z_1, \dots, Z_n)^\top$.

Let $\boldsymbol{\sigma}$ be the square root of $\mathbf{V}$. Then, $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{Z}$ is a Gaussian vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. *i.e.*, $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \mathbf{V})$.

Proof. $\boldsymbol{X}$ is a Gaussian vector as a linear transformation of a Gaussian vector. We have

$$\begin{aligned} \mathbb{E}[\boldsymbol{X}] &= \boldsymbol{\mu} \\ \operatorname{Var}(\boldsymbol{X}) &= \mathbb{E}\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\big] \\ &= \mathbb{E}\big[(\boldsymbol{\sigma}\boldsymbol{Z})(\boldsymbol{\sigma}\boldsymbol{Z})^\top\big] \\ &= \boldsymbol{\sigma}\mathbb{E}\big[\boldsymbol{Z}\boldsymbol{Z}^\top\big]\boldsymbol{\sigma}^\top \\ &= \boldsymbol{\sigma}\mathbf{I}_n\boldsymbol{\sigma}^\top = \boldsymbol{\sigma}^2 = \mathbf{V}. \end{aligned}$$

4.9.3 Density of a Gaussian Vector

Let $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \mathbf{V})$. We want to find the density of $\boldsymbol{X}$.

[In the $n = 1$ case, we have $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.]

We shall consider two cases

- $\mathbf{V}$ is positive definite, *i.e.* $\lambda_1, \dots, \lambda_n > 0$. We can write

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{Z} \quad \text{where } \boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{I}_n).$$

  Note that $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{z}$ gives $\boldsymbol{z} = \boldsymbol{\sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$. Hence,

$$\begin{aligned} f_X(x) &= f_Z\left(\sigma^{-1}(x-\mu)\right) \cdot \left|\det \sigma^{-1}\right| \\ &= f_Z(z_1, \ldots, z_n) \cdot \det \sigma^{-1} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{|z|^2}{2}\right) \cdot \det \sigma^{-1} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \det \sigma} \exp\left(-\frac{(x-\mu)^\top \cdot \left(\sigma^{-1}\right)^\top \cdot \sigma^{-1} \cdot (x-\mu)}{2}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\frac{(x-\mu)^\top \cdot V^{-1} \cdot (x-\mu)}{2}\right). \end{aligned}$$

Lecture 21 · 2026-03-11

- V is non-negative definite, and $\exists i$ such that $\lambda_i = 0$.

  By an orthogonal change of basis, we could assume that

  $$V = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix} \quad \text{where } U \text{ is a positive definite matrix of size } m \times m \text{ for some } m < n.$$

  Let

  $$\mu = \begin{pmatrix} \lambda \\ \nu \end{pmatrix} \quad \text{where } \lambda \in \mathbb{R}^m \text{ and } \nu \in \mathbb{R}^{n-m}.$$

  We can then write

  $$X = \begin{pmatrix} Y \\ \nu \end{pmatrix} \text{ where } Y \sim N(\lambda, U)$$

  $$f_Y(y) = \frac{1}{\sqrt{(2\pi)^m \det U}} \exp\left(-\frac{(y-\lambda)^\top \cdot U^{-1} \cdot (y-\lambda)}{2}\right).$$

Proposition 4.47

Let $X = (X_1, \ldots, X_n)$ be a Gaussian vector. Let $\mu = \mathbb{E}[X]$ and $V = \mathrm{Var}(X)$.

If $X_1, \ldots, X_n$ are independent, then V is a diagonal matrix.

Proof. Since $V_{ij} = \mathrm{Cov}\left(X_i, X_j\right) = 0$ for $i \neq j$, the matrix V is diagonal due to independence.

Proposition 4.48

Let $X = (X_1, \ldots, X_n)$ be a Gaussian vector. If V is a diagonal matrix and strictly positive definite, then $X_1, \ldots, X_n$ are independent.

Proof 1. We have

$$\mathbf{V} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{pmatrix} \quad \text{where } \lambda_1, \dots, \lambda_n > 0.$$

Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$. We have

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{\sqrt{(2\pi)^n \det \mathbf{V}}} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{\mu})^\top \cdot \mathbf{V}^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})}{2} \right) \\ &= \frac{1}{\sqrt{(2\pi)^n \det \mathbf{V}}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\lambda_i} \right). \end{aligned}$$

Since $f_{\boldsymbol{X}}$ factorises, by Theorem 4.23 (2), $X_1, \dots, X_n$ are independent, and $X_i \sim N(\mu_i, \lambda_i)$.

Proof 2. Consider the MGF of $\boldsymbol{X}$. We have

$$m(\theta) = \mathbb{E}\left[e^{\theta^\top X} \right] = \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} + \frac{1}{2} \boldsymbol{\theta}^\top \mathbf{V} \boldsymbol{\theta} \right).$$

since we have

$$\boldsymbol{\theta}^\top \boldsymbol{X} \sim N(\boldsymbol{\theta}^\top \boldsymbol{\mu}, \boldsymbol{\theta}^\top \mathbf{V} \boldsymbol{\theta}).$$

Hence

$$\begin{aligned} m(\theta) &= \exp\left(\sum_{i=1}^{n} \theta_i \mu_i + \sum_{i=1}^{n} \frac{\theta_i^2 \lambda_i}{2} \right) \\ &= \prod_{i=1}^{n} \exp\left(\theta_i \mu_i + \frac{\theta_i^2 \lambda_i}{2} \right) \end{aligned}$$

Since $m(\theta)$ factorises into a product of functions of θ_i's, by Theorem 4.32, $X_1, \dots, X_n$ are independent, and $X_i \sim N(\mu_i, \lambda_i)$.

Therefore, we can conclude that if $(X_1, \dots, X_n)$ is a Gaussian vector, then $X_1, \dots, X_n$ are independent iff $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$.

4.9.4 Bivariate Gaussian Distribution

Definition 4.49

Let $\boldsymbol{X} = (X_1, X_2)$ be a Gaussian vector in $\mathbb{R}^2$. Let $\mu_k = \mathbb{E}[X_k]$, $\sigma_k^2 = \operatorname{Var}(X_k)$ and

$$\rho = \operatorname{Corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1) \operatorname{Var}(X_2)}}.$$

Then $\boldsymbol{X}$ is called a **bivariate Gaussian vector** with parameters μ_1, μ_2, σ_1^2 and σ_2^2.

Proposition 4.50

$\rho \in [-1, 1]$.

Proof. By Cauchy-Schwarz inequality, the result follows.

Note that we can write the covariance matrix of $\boldsymbol{X}$ as

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Proposition 4.51

Let $\sigma_1, \sigma_2 > 0$ and $\rho \in [-1, 1]$. Then the matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

is non-negative definite.

Proof. Consider any $\boldsymbol{u} = (u_1, u_2)^\top \in \mathbb{R}^2$. Then

$$\begin{aligned} \boldsymbol{u}^\top \mathbf{V} \boldsymbol{u} &= (1-\rho)\left(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2\right) + \rho(\sigma_1 u_1 + \sigma_2 u_2)^2 \\ &= (1+\rho)\left(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2\right) - \rho(\sigma_1 u_1 - \sigma_2 u_2)^2 \end{aligned}$$

- If $\rho \in [-1, 0]$, the second line gives $\boldsymbol{u}^\top \mathbf{V} \boldsymbol{u} \geqslant 0$.
- If $\rho \in [0, 1]$, the first line gives $\boldsymbol{u}^\top \mathbf{V} \boldsymbol{u} \geqslant 0$.

Now, consider $\mathbb{E}[X_2 \mid X_1]$. We can write

$$X_2 = X_2 - aX_1 + aX_1 \quad \text{for any } a \in \mathbb{R}.$$

Let $Y = X_2 - aX_1$. Then $\binom{X_1}{Y}$ is a Gaussian vector since

$$\begin{pmatrix} X_1 \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Note that

$$\begin{aligned} \operatorname{Cov}(X_1, Y) &= \operatorname{Cov}(X_1, X_2 - aX_1) \\ &= \operatorname{Cov}(X_1, X_2) - a \operatorname{Var}(X_1). \end{aligned}$$

Taking $a = \frac{\operatorname{Cov}(X_1, X_2)}{\operatorname{Var}(X_1)}$ gives $\operatorname{Cov}(X_1, Y) = 0$. Hence, X_1 and $Y = X_2 - aX_1$ are independent. We have

$$\begin{aligned} \mathbb{E}[X_2 \mid X_1] &= \mathbb{E}[X_2 - aX_1 \mid X_1] + \mathbb{E}[aX_1 \mid X_1] \\ &= \mathbb{E}[X_2 - aX_1] + aX_1 \\ &= \mu_2 - a\mu_1 + aX_1 \\ &= \mu_2 + \left(\rho\frac{\sigma_2}{\sigma_1}\right)(X_1 - \mu_1). \end{aligned}$$

5 Convergence Results and Limit Theorems

5.1 Convergence Results

Definition 5.1 (Convergence in Probability)

A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges to a random variable X **in probability** and we write

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$$

if as $n \to \infty$,

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X_n - X| > \varepsilon) \to 0.$$

Theorem 5.2 (Weak Law of Large Numbers)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X_i] < \infty$. Let $S_n = X_1 + \ldots + X_n$. Then

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\mathbb{P}} \mu.$$

i.e.

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

This is called the **weak law of large numbers** (WLLN).

Lecture 22 · 2026-03-13

Proof. Assume that $\sigma^2 = \mathrm{Var}(X_1) < \infty$. We need to show that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

We have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) &= \mathbb{P}(|S_n - n\mu| > n\varepsilon) \\ &\leqslant \frac{\mathbb{E}\left[|S_n - n\mu|^2\right]}{n^2\varepsilon^2} \quad \text{by Chebyshev's} \\ &= \frac{\mathrm{Var}(S_n)}{n^2\varepsilon^2}.\end{aligned}$$

By definition, we have $\mathrm{Var}(S_n) = n\sigma^2$. So

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leqslant \frac{\sigma^2 n}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

Definition 5.3 (Convergence Almost Surely)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables and X be a random variable. We say that X_n converges to X **almost surely** (a.s.) or **with probability 1** if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

We write $X_n \to X$ as $n \to \infty$ almost surely.

Remark. The above statement, more precisely, means that

$$\exists A \in \mathcal{F} \text{ with } \mathbb{P}(A) = 1, \forall \omega \in A, \forall \varepsilon > 0, \exists n_0, \forall n \geqslant n_0 : \left|X_n(\omega) - X(\omega)\right| \leqslant \varepsilon$$

where $X_n : \Omega \to \mathbb{R}$ and $X : \Omega \to \mathbb{R}$.

To compare with Definition 5.1, this is equivalently saying that

$$\mathbb{P}(\forall \varepsilon > 0, \exists n_0, \forall n \geqslant n_0 : \left|X_n - X\right| \leqslant \varepsilon) = 1.$$

Note that the quantifiers are now inside $\mathbb{P}$.

[It can be clearer to see the difference between **convergence in probability** and **convergence almost surely** without any shorthands. Let the underlying probability space be $(\omega, \mathcal{F}, \mathbb{P})$. Let X and the sequence $X_1, X_2, \ldots$ be random variables, where $X, X_i : \Omega \to \mathbb{R}$.

Convergence in probability states that $\forall \varepsilon > 0$,

$$\begin{aligned} & \lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega : \left|X_n(\omega) - X(\omega)\right| > \varepsilon\}) = 0 \\ \Leftrightarrow \quad & \forall \varepsilon > 0, \forall \delta > 0, \exists n_0 \in \mathbb{N}, \forall n \geqslant n_0, \mathbb{P}(\{\omega \in \Omega : \left|X_n(\omega) - X(\omega)\right| > \varepsilon\}) < \delta. \end{aligned}$$

Convergence almost surely states that

$$\begin{aligned} & \mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1 \\ \Leftrightarrow \quad & \mathbb{P}(\{\omega \in \Omega : \forall \varepsilon > 0, \exists n_0 \in \mathbb{N}, \forall n \leqslant n_0, \left|X_n(\omega) - X(\omega)\right| < \varepsilon\}) = 1. \end{aligned}$$

i.e. **convergence in probability** cares about that as $n \to \infty$, (at each snapshot of n big enough) the percentage of the population Ω that behaves badly (away from $X(\omega)$) shrinks to zero; whereas **convergence almost surely** cares about that the percentage of the population that converges to $X(\omega)$ as $n \to \infty$ (without leaving there) is 1. This is a much stronger condition. For a sequence to converge almost surely, the individuals can't keep jumping away from the target infinitely often, but they could for convergence in probability.

This is indeed a challenging topic. See Wikipedia, Maths Stack Exchange for further explanation.]

Proposition 5.4

If $X_n \to 0$ as $n \to \infty$ almost surely, then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$.

Proof. We need to show that $\forall \varepsilon > 0, \mathbb{P}(\left|X_n\right| \leqslant \varepsilon) \to 1$ as $n \to \infty$. We have

$$\mathbb{P}(|X_n| \leqslant \varepsilon) \geqslant \mathbb{P}\left(\underbrace{\bigcap_{m=n}^{\infty} |X_m| \leqslant \varepsilon}_{A_n}\right).$$

Note that $A_n \subseteq A_{n+1}$. Hence

$$\mathbb{P}(A_n) \xrightarrow[n\to\infty]{\text{increasing}} \mathbb{P}\left(\bigcup_n A_n\right) = \mathbb{P}\left(\bigcup_A \bigcap_{m=n}^{\infty} \{|X_m| \leqslant \varepsilon\}\right).$$

So

$$\lim_{n\to\infty} \mathbb{P}(|X_n| \leqslant \varepsilon) \geqslant \mathbb{P}(\forall \varepsilon > 0, |X_m| \leqslant \varepsilon \text{ for all } m \text{ sufficiently large}).$$

By assumption, the RHS is exactly 1.

Remark. In general, almost sure convergence implies convergence in probability, but not the other way around.

Theorem 5.5 (Strong Law of Large Numbers)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X_i] < \infty$. Let $S_n = X_1 + \ldots + X_n$. Then

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{almost surely}} \mu.$$

This is called the **strong law of large numbers** (SLLN).

Proof. [Non-examinable.]

Assume that $E\left[X_1^4\right] < \infty$. Set $Y_i = X_i - \mu$. Then $\mathbb{E}[Y_i] = 0$ and

$$\mathbb{E}\left[Y_1^4\right] = \mathbb{E}\left[(X_1 - \mu)^4\right] \leqslant 2^4\left(\mathbb{E}\left[X_1^4\right] + \mu^4\right) < \infty.$$

Set $S_n = \sum_{i=1}^n Y_i$. We need to show that $\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = 0\right) = 1$.

The goal is to show that $\mathbb{P}\left(\sum_{n=1}^{\infty} \left(\frac{S_n}{n}\right)^4 < \infty\right) = 1$ since this will imply that with probability 1, $\frac{S_n}{n} \to 0$ as $n \to \infty$.

It suffices to show that $\mathbb{E}\left[\sum_{n=1}^{\infty} \left(\frac{S_n}{n}\right)^4\right] < \infty$. We have

$$\mathbb{E}\left[\sum \left(\frac{S_n}{n}\right)^4\right] = \sum_n \frac{1}{n^4}\mathbb{E}\left[S_n^4\right].$$

By definition, we have

$$\mathbb{E}\left[S_n^4\right] = \mathbb{E}\left[(Y_1 + \ldots + Y_n)^4\right] = \mathbb{E}\left[\sum_{i=1}^n Y_i^4 + 6 \sum_{1\leqslant i<j\leqslant n} Y_i^2 Y_j^2 + R\right]$$

where R is a sum of terms of the form

$$Y_i^3 Y_j, Y_i^2 Y_j Y_k, Y_i Y_j Y_k Y_\ell$$

for distinct i, j, k, ℓ. By independence and the fact that $\mathbb{E}[Y_i] = 0$, we have $\mathbb{E}[R] = 0$. Moreover,

$$\mathbb{E}\left[Y_i^2 Y_j^2\right] = \left(\mathbb{E}\left[Y_i^2\right]\right)^2 \leqslant \mathbb{E}\left[Y_i^4\right] < \infty.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[S_n^4\right] &= n\mathbb{E}\left[Y_1^4\right] + \frac{6(n)(n-1)}{2}\mathbb{E}\left[Y_1^4\right] \\ &\leqslant 3n^2\mathbb{E}\left[Y_1^4\right].\end{aligned}$$

So

$$\sum \frac{\mathbb{E}\left[S_n^4\right]}{n^4} \leqslant \sum \frac{3\mathbb{E}\left[Y_1^4\right]}{n^2} < \infty.$$

5.2 Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X_1] < \infty$ and variance $\sigma^2 = \mathrm{Var}(X_1) < \infty$. Set $S_n = X_1 + \ldots + X_n$. By SLLN 5.5, we expect that

$$S_n \approx n\mu \quad \text{for large } n.$$

Note that

$$\mathrm{Var}(S_n - n\mu) = n\sigma^2, \quad \mathbb{E}[S_n - n\mu] = 0.$$

Hence $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ has expectation 0 and variance 1.

Also, we have

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\frac{S_n}{n} - \mu}{\sqrt{\mathrm{Var}\left(\frac{S_n}{n} - \mu\right)}}.$$

So, if $X_1, X_2, \ldots$ are i.i.d. with $X_i \sim N\left(\mu, \sigma^2\right)$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \sim N(0, 1).$$

The normal distribution is universal as it appears as the limit of the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ no matter what the distribution of X_i is, as long as X_i are i.i.d. with mean μ and variance σ^2. This is the content of the central limit theorem (CLT).

Definition 5.6 (Convergence in Distribution)

A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges to X in distribution as $n \to \infty$ and we write

$$X_n \xrightarrow[n\to\infty]{(d)} X \quad \text{if} \quad \mathbb{P}(X_n \leqslant x) = F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x) = \mathbb{P}(X \leqslant x)$$

for all x where F_X is continuous.

Lecture 23 · 2026-03-16

Proposition 5.7

If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then $X_n \xrightarrow[n\to\infty]{(d)} X$.

Proof. Let x be a continuity point of F_X. We need to show that $F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x)$. Let $\varepsilon > 0$.

We have

$$\{X \leqslant x - \varepsilon\} \subseteq \{X_n \leqslant x\} \cup \{|X_n - X| > \varepsilon\}$$
$$\{X_n \leqslant x\} \subseteq \{X \leqslant x + \varepsilon\} \cup \{|X_n - X| > \varepsilon\}.$$

So we have

$$\mathbb{P}(X \leqslant x - \varepsilon) \leqslant \mathbb{P}(X_n \leqslant x) + \mathbb{P}(|X_n - X| > \varepsilon)$$
$$F_X(x - \varepsilon) \leqslant \liminf_{n\to\infty} \mathbb{P}(X_n \leqslant x)$$

Also,

$$\mathbb{P}(X_n \leqslant x) \leqslant \mathbb{P}(X \leqslant x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)$$
$$\limsup_{n\to\infty} F_{X_n}(x) \leqslant F_X(x + \varepsilon).$$

Hence

$$F_X(x - \varepsilon) \leqslant \liminf_{n\to\infty} F_{X_n}(x) \leqslant \limsup_{n\to\infty} F_{X_n}(x) \leqslant F_X(x + \varepsilon).$$

Therefore, letting $\varepsilon \to 0$, we have

$$F_X(x) \leqslant \liminf_{n\to\infty} F_{X_n}(x) \leqslant \limsup_{n\to\infty} F_{X_n}(x) \leqslant F_X(x).$$

So $F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x)$ as required.

Theorem 5.8 (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X_i] < \infty$ and variance $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Set $S_n = X_1 + \ldots + X_n$. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{(d)} N(0,1) = Z.$$

i.e. $\forall x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) \xrightarrow[n\to\infty]{} \int_{-\infty}^{x} \frac{\exp\left(-\frac{t^y}{2}\right)}{\sqrt{2\pi}}\,\mathrm{d}y = \Phi(x).$$

Proof. We will need the following continuity property for moment generating functions.

Theorem 5.9 (Continuity Property for MGFs)

Suppose (X_n) are random variables with $m_n(\theta) = \mathbb{E}\left[e^{\theta X_n}\right]$ for $\theta \in \mathbb{R}$ and X is a random variable with $m(\theta) = \mathbb{E}\left[e^{\theta X}\right]$ for $\theta \in \mathbb{R}$. Assume $m(\theta) < \infty$ for some $\theta \neq 0$.

If $m_n(\theta) \to m(\theta)$ as $n \to \infty$ for all θ in $\mathbb{R}$, then

$$X_n \xrightarrow[n\to\infty]{(d)} X.$$

The proof is beyond the scope of this course.

Consider $Y_i = \frac{X_i - \mu}{\sigma}$. Then

$$\mathbb{E}[Y_i] = 0, \quad \mathrm{Var}(Y_i) = 1.$$

It is enough to prove the theorem for a sequence $X_1, X_2, \ldots$, i.i.d. with mean 0 and variance 1. Set $S_n = X_1 + \ldots + X_n$. We need to show that

$$\frac{S_n}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} Z.$$

Assume that $\exists \delta > 0$ such that $\mathbb{E}\left[e^{\delta X_1}\right] + \mathbb{E}\left[-e^{\delta X_1}\right] < \infty$.

Set $m(\theta) = \mathbb{E}\left[e^{\theta X_1}\right]$. By the continuity property for MGFs, it suffices to show that

$$\mathbb{E}\left[e^{\theta \frac{S_n}{\sqrt{n}}}\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[e^{\theta Z}\right] = \exp\left(\frac{\theta^2}{2}\right).$$

Note that

$$\mathbb{E}\left[e^{\theta \frac{S_n}{\sqrt{n}}}\right] = \left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n$$

We want to show that

$$\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n \xrightarrow[n\to\infty]{} \exp\left(\frac{\theta^2}{2}\right) \quad \forall \theta \in \mathbb{R}.$$

We have

$$m\left(\frac{\theta}{\sqrt{n}}\right) = \mathbb{E}\left[e^{\frac{\theta X_1}{\sqrt{n}}}\right] = 1 + \frac{\theta^2}{2n} + \mathbb{E}\left[\sum_{k\geqslant 3} \frac{(\theta X_1)^k}{\left(\sqrt{n}\right)^k (k!)}\right].$$

Claim.

$$\left|\mathbb{E}\left[\sum_{k\geqslant 3} \frac{(\theta X_1)^k}{k!}\right]\right| = o\left(|\theta|^2\right) \quad \text{as } n \to \infty.$$

Proof. Let $|\theta| < \frac{\delta}{2}$.

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{k\geqslant 3}\frac{(\theta X_1)^k}{k!}\right]\right| &\leqslant \mathbb{E}\left[\sum_{k\geqslant 3}|\theta|^k\frac{|X_1|^k}{k!}\right]\\ &= \mathbb{E}\left[|\theta X_1|^3\sum_{k\geqslant 0}\frac{|\theta X_1|^k}{(k+3)!}\right]\\ &\leqslant \mathbb{E}\left[|\theta X_1|^3 e^{|\theta X_1|}\right]\\ &\leqslant \mathbb{E}\left[|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|}\right].\end{aligned}$$

However,

$$|\theta X_1|^3 = |\theta|^3\frac{\left|\frac{\delta}{2}X_1\right|^3}{3!}\cdot\frac{3!}{\left(\frac{\delta}{2}\right)^3} \leqslant C\cdot|\theta|^3 e^{\frac{\delta}{2}|X_1|}$$

where $C = 3!\cdot\frac{2^3}{\delta^3}$. So

$$\begin{aligned}\mathbb{E}\left[|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|}\right] &\leqslant C\cdot|\theta|^3\mathbb{E}\left[e^{\delta|X_1|}\right]\\ &\leqslant C\cdot|\theta|^3\left(\mathbb{E}\left[e^{\delta X_1}\right]+\mathbb{E}\left[-e^{\delta X_1}\right]\right) < \infty.\end{aligned}$$

So

$$\left|\mathbb{E}\left[\sum_{k\geqslant 3}\frac{(\theta X_1)^k}{k!}\right]\right| \leqslant C'\cdot|\theta|^3 = o\left(|\theta|^2\right) \quad \text{as } \theta\to 0$$

where $C' = C\left(\mathbb{E}\left[e^{\delta X_1}\right]+\mathbb{E}\left[-e^{\delta X_1}\right]\right)$.

Then we can conclude, because

$$m\left(\frac{\theta}{\sqrt{n}}\right) = 1+\frac{\theta^2}{2n}+o\left(\frac{\theta^2}{n}\right) \quad \text{as } n\to\infty,$$

and hence

$$\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n \to \exp\left(\frac{\theta^2}{2}\right) \quad \text{as } n\to\infty.$$

Corollary 5.10

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X_i] < \infty$ and variance $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Set $S_n = X_1 + \ldots + X_n$. Then

$$S_n \approx N\left(n\mu, n\sigma^2\right) \quad \text{for large } n.$$

Example 5.11

- Suppose $S_n \sim \text{Bin}(n, p)$. Then $S_n = X_1 + \ldots + X_n$ where X_i are i.i.d. with $X_i \sim \text{Ber}(p)$. So

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow[n\to\infty]{(d)} N(0,1).$$

  Therefore, for large n,

$$S_n \approx N(np, np(1-p)).$$

- Suppose $S_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ with $\lambda > 0$. Then

$$\mathbb{P}(S_n = x) \xrightarrow[n\to\infty]{} \mathbb{P}(\text{Poi}(\lambda) = x) \quad \forall x \in \mathbb{N}.$$

- We can approximate Poisson distribution with normal distribution. Suppose $S_n \sim \text{Poi}(n)$ with $n > 0$. Then $S_n = X_1 + \ldots + X_n$ where X_i are i.i.d. with $X_i \sim \text{Poi}(1)$. So

$$\frac{S_n - n}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} N(0,1).$$

  Therefore, for large n,

$$S_n \approx N(n, n).$$

Lecture 24 · 2026-03-18

5.2.1 Sampling Error via the CLT

Suppose we hold a referendum and a proportion p of individuals is inclined to write "Yes". [Every individual writes "Yes" with probability p and "No" with probability $1 - p$.] We wish to estimate p.

Sample N individuals at random (for a large N) and record their votes. Let X_i be 1 if the i-th individual voted "Yes" and 0 otherwise.

Let $S_N = X_1 + \ldots + X_N$ be the number of "Yes" votes. Then by SLLN 5.5,

$$\hat{p}_N := \frac{S_N}{N} \xrightarrow[N\to\infty]{\text{almost surely}} p$$

We have $S_N \sim \text{Bin}(N, p)$.

We need to testimate p with an accuracy of $\pm 4\%$ with probability $\geqslant 0.99$. We now wish to consider how large should N be.

By Central Limit Theorem 5.8,

$$S_N \approx Np + \sqrt{Np(1-p)}Z$$

where $Z \sim N(0,1)$ for large N. Then we want to find N such that

$$\mathbb{P}(|\hat{p}_N - p| \leqslant 0.04) \leqslant 0.01.$$

Note that for large N,

$$\hat{p}_N = \frac{S_N}{N} \approx p + \sqrt{\frac{p(1-p)}{N}} Z.$$

Hence

$$|\hat{p}_N - p| \approx \sqrt{\frac{p(1-p)}{N}} |Z|.$$

Since we have $\mathbb{P}(|Z| \geqslant z) = 2(1 - \Phi(z))$ and that $\mathbb{P}(|Z| \geqslant 2.58) = 0.01$ by standard tables, we requrie

$$\frac{\sqrt{N}}{\sqrt{p(1-p)}} \cdot 0.04 \geqslant 2.58$$

Since $\sqrt{p(1-p)} \leqslant \frac{1}{2}$, taking $N = 1040$ is sufficient.

5.3 Simulation of Random Variables

Computers can generate random numbers between 0 to 1. We would like to extend this to general probability distributions beyond uniform distributions.

Example 5.12

Suppose X is a random variable with probability distribution function $F(x) = \mathbb{P}(X \leqslant x)$. Suppose that we know how to simulate $U \sim U[0,1]$.

- Assume that F is 1-1.

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

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

- If F is not 1-1, then define its generalised inverse $G(u) = \inf\{x \in \mathbb{R} : u \leqslant F(x)\}$. Then let $X = F^{-1}(U)$. We can show that $\mathbb{P}(X \leqslant x) = F(x)$ as well.

  > ***Claim.*** $G(u) \leqslant x$ iff $u \leqslant F(x)$.

  > ***Proof.*** [$\Leftarrow$] This follows by definition.
  >
  > [$\Rightarrow$] $\exists (x_n) \geqslant G(u)$ such that x_n decreases to $G(u)$ and $u \leqslant F(x_n)$. By the right continuity of F,
  >
  > $$\lim_{n\to\infty} F(x_n) = F(G(u)).$$
  >
  > So $u \leqslant F(G(u))$. Therefore, if $G(u) \leqslant x$, then $u \leqslant F(G(u)) \leqslant F(x)$ becuase F is increasing.

  Let $X = G(U)$. Then

  $$\mathbb{P}(X \leqslant x) = \mathbb{P}(G(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x).$$

5.3.1 Box-Muller Transform

We want to simulate X, Y independent with $X, Y \sim N(0,1)$ using two independent $U[0,1]$. Since

$$F(x) = \int_{-\infty}^{x} \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t,$$

there is no closed form solution. However, recall Example 4.29.

Let $U, V \sim U[0,1]$ be independent. Then let

$$\Theta = 2\pi U \sim U[0, 2\pi]$$
$$R = \sqrt{-2\log V}$$

We get

$$\mathbb{P}(R \geqslant r) = \mathbb{P}\left(\sqrt{-2\log V} \geqslant r\right) = \mathbb{P}\left(V \leqslant \mathrm{e}^{-\frac{r^2}{2}}\right) = \mathrm{e}^{-\frac{r^2}{2}}$$

and hence R has the correct density. Using the transformation in Example 4.29, we get X, Y:

$$X = R\cos\Theta = \sqrt{-2\log V}\cos(2\pi U)$$
$$Y = R\sin\Theta = \sqrt{-2\log V}\sin(2\pi U).$$

5.4 Rejection Sampling

Suppose we have a random variable X with density

$$f(\boldsymbol{x}) = \frac{\mathbb{1}(\boldsymbol{x} \in A)}{|A|}$$

where $|A|$ is the volume of $A \subseteq [0,1]^d$.

Let $(U_{k,n})_{k=1,\ldots,d,n\in\mathbb{N}}$ be i.i.d. with $U_{k,n} \sim U[0,1]$.

Set $\boldsymbol{U_n} = (U_{1,n}, \ldots, U_{d,n}) \sim U\left([0,1]^d\right)$. Let

$$N = \min\{n : \boldsymbol{U_n} \in A\}, \quad \boldsymbol{X} = \boldsymbol{U}_N.$$

We want to show that $\forall B \subseteq [0,1]^d$,

$$\int_B f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{P}(\boldsymbol{X} \in B) = \frac{|B \cap A|}{|A|}.$$

Note that

$$\begin{aligned}
\mathbb{P}(\boldsymbol{X} \in B) = \mathbb{P}(\boldsymbol{U}_N \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(\boldsymbol{U_n} \in B, N \in n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(\boldsymbol{U_n} \in B, \boldsymbol{U_n} \in A, \boldsymbol{U_{n-1}} \notin A, \ldots, \boldsymbol{U_1} \notin A) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(\boldsymbol{U_n} \in B \cap A)\mathbb{P}(\boldsymbol{U_1} \notin A)^{n-1} \\
&= \sum_{n=1}^{\infty} |B \cap A|(1 - |A|)^{n-1} = \frac{|B \cap A|}{|A|}.
\end{aligned}$$

END OF DOCUMENT ■