

MATH 394: Probability I

Supplementary Notes
Department of Mathematics
University of Washington

Summer 2026

ARMAN JAHANGIRI

1 Continuity of Probability

Motivation

In probability, we often describe complicated events as limits of simpler events. For example, suppose we have events

$$A_1, A_2, \dots, A_n, \dots$$

and we want to understand the probability of an event obtained by taking an infinite union or an infinite intersection. The key fact is that the probability function $\mathbb{P}(\cdot)$ behaves continuously under limits of certain sequences of events, in a precise mathematical sense. This is called the **continuity of probability**, which we will prove here. Before that, we revisit the probability axioms.

Probability Space

A *probability space* is a triple

$$(S, \mathcal{F}, \mathbb{P}),$$

where

(1) **Sample space.**

S is the set of all possible outcomes of the experiment.

(2) **Sigma-algebra.**

$\mathcal{F}$ is a collection of subsets of S , called *events*, satisfying:

(a)

$$S \in \mathcal{F}.$$

(b) If $A \in \mathcal{F}$, then

$$A^c \in \mathcal{F}.$$

(c) If

$$A_1, A_2, \dots, A_n, \dots \in \mathcal{F},$$

then

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

A collection satisfying these three properties is called a σ -algebra (or *sigma-field*).

(3) **Probability measure.**

$\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a function that assigns a probability to every event in $\mathcal{F}$. Such a function is called a *probability measure* and satisfies the following axioms.

(A1) **Nonnegativity:**

$$\mathbb{P}(A) \geq 0, \quad \text{for every } A \in \mathcal{F}.$$

(A2) **Normalization:**

$$\mathbb{P}(S) = 1.$$

(A3) **Countable additivity:**

If

$$A_1, A_2, \dots, A_n, \dots$$

are pairwise disjoint events, then

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

Increasing and Decreasing Sequences of Events

Definition 1 (Increasing sequence of events). A sequence of events $(A_n)_{n \geq 1}$ is called **increasing** if

$$A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots.$$

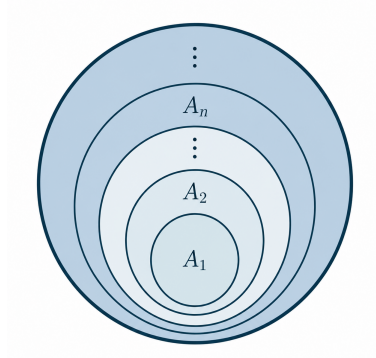


Figure 1: An increasing sequence of events, $A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots$, whose limit is $\bigcup_{n=1}^{\infty} A_n$.

In this case, we write

$$A_n \uparrow \bigcup_{n=1}^{\infty} A_n,$$

and we define

$$\lim_{n \rightarrow \infty} A_n := \bigcup_{n=1}^{\infty} A_n.$$

So $A_n \uparrow \lim_{n \rightarrow \infty} A_n$ means that the events are getting larger and their limiting event is the union.

Definition 2 (Decreasing sequence of events). A sequence of events $(A_n)_{n \geq 1}$ is called **decreasing** if

$$A_1 \supseteq A_2 \supseteq \dots \supseteq A_n \supseteq \dots.$$

In this case, we write

$$A_n \downarrow \bigcap_{n=1}^{\infty} A_n,$$

where

$$\lim_{n \rightarrow \infty} A_n := \bigcap_{n=1}^{\infty} A_n.$$

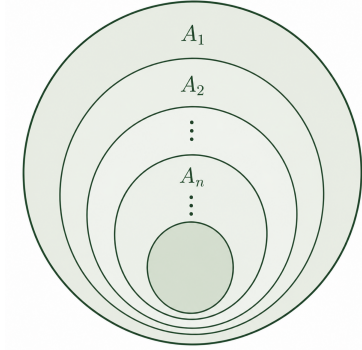


Figure 2: A decreasing sequence of events, $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$, whose limit is $\bigcap_{n=1}^{\infty} A_n$.

So $A_n \downarrow \bigcap_{n=1}^{\infty} A_n$ means that the events are getting smaller and their limiting event is the intersection.

Continuity of the probability function from Below

Theorem 1 (Continuity from below). *Suppose $(A_n)_{n \geq 1}$ is a sequence of events in $\mathcal{F}$, and*

$$A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$$

Then

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

Equivalently, if $A_n \uparrow A$, then

$$\mathbb{P}(A_n) \uparrow \mathbb{P}(A).$$

Proof. Define disjoint pieces

$$B_1 = A_1, \quad B_n = A_n \setminus A_{n-1} \quad \text{for } n \geq 2.$$

These are disjoint as for any $m < n$, we have

$$B_n \cap B_m = (A_n - A_{n-1}) \cap (A_m - A_{m-1}) = A_n \cap A_{n-1}^c \cap A_m \cap A_{m-1}^c$$

Since $m < n$, we have $A_m \subseteq A_n$, and $A_{m-1} \subseteq A_{n-1}$, and therefore, $A_{n-1}^c \subseteq A_{m-1}^c$. This implies that $A_n \cap A_m = A_m$ and $A_{n-1}^c \cap A_{m-1}^c = A_{n-1}^c$. Now, since $n-1 \leq m$, $A_{n-1} \subseteq A_m$ and therefore, $A_m \cap A_{n-1}^c = \emptyset$. Hence, we have shown

$$A_n \cap A_{n-1}^c \cap A_m \cap A_{m-1}^c = A_m \cap A_{n-1}^c = \emptyset$$

So $B_n \cap B_m = \emptyset$.

Next, we claim (please show) that for every $n \geq 1$,

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

Also,

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n.$$

Therefore, by countable additivity,

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

On the other hand, for each fixed n ,

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

Taking the limit as $n \rightarrow \infty$, we get

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \mathbb{P}(B_k) = \sum_{k=1}^{\infty} \mathbb{P}(B_k).$$

Thus,

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right).$$

□

Remark 1 (Analogy with left-continuity of functions). Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous from the left at a point a . If

$$a_n \uparrow a,$$

meaning

$$a_1 \leq a_2 \leq \cdots \leq a_n \leq \cdots \quad \text{and} \quad a_n \rightarrow a,$$

then

$$f(a_n) \rightarrow f(a).$$

Equivalently,

$$\lim_{n \rightarrow \infty} f(a_n) = f\left(\lim_{n \rightarrow \infty} a_n\right) = f(a).$$

Continuity from below for probability says something very similar. If

$$A_n \uparrow A,$$

then

$$\mathbb{P}(A_n) \rightarrow \mathbb{P}(A).$$

Equivalently,

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\lim_{n \rightarrow \infty} A_n\right)$$

In this sense, probability is continuous along increasing sequences of events. The role of “approaching from the left” is played by the inclusions

$$A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots \subseteq \bigcup_{i=1}^{\infty} A_i.$$

Continuity of the probability function from Above

Theorem 2 (Continuity from above). *Suppose $(A_n)_{n \geq 1}$ is a sequence of events in $\mathcal{F}$, and*

$$A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots .$$

Then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

Equivalently, if $A_n \downarrow A$, then

$$\mathbb{P}(A_n) \downarrow \mathbb{P}(A).$$

Proof. Let

$$A = \bigcap_{n=1}^{\infty} A_n.$$

Since $A_n \downarrow A$, the complements satisfy

$$A_1^c \subseteq A_2^c \subseteq \cdots \subseteq A_n^c \subseteq \cdots .$$

Also, by De Morgan's law,

$$\bigcup_{n=1}^{\infty} A_n^c = \left(\bigcap_{n=1}^{\infty} A_n\right)^c = A^c.$$

By continuity from below,

$$\mathbb{P}(A^c) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^c\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n^c).$$

Using complements,

$$\mathbb{P}(A_n^c) = 1 - \mathbb{P}(A_n),$$

and

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

Therefore,

$$1 - \mathbb{P}(A) = \lim_{n \rightarrow \infty} (1 - \mathbb{P}(A_n)).$$

Hence,

$$\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

□

Remark 2 (Analogy with right-continuity of functions). Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous from the right at a point a . If

$$a_n \downarrow a,$$

meaning

$$a_1 \geq a_2 \geq \cdots \geq a_n \geq \cdots \quad \text{and} \quad a_n \rightarrow a,$$

then

$$f(a_n) \rightarrow f(a).$$

Equivalently,

$$\lim_{n \rightarrow \infty} f(a_n) = f\left(\lim_{n \rightarrow \infty} a_n\right) = f(a).$$

Continuity from above for probability says the analogous statement for decreasing sequences of events. If

$$A_n \downarrow A,$$

then

$$\mathbb{P}(A_n) \rightarrow \mathbb{P}(A).$$

Here, the role of “approaching from the right” is played by

$$A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots \supseteq A.$$

“latex

Properties of a Cumulative Distribution Function

Let X be a random variable on the probability space

$$(S, \mathcal{F}, \mathbb{P}).$$

The **cumulative distribution function**, or **CDF**, of X is the function

$$F : \mathbb{R} \rightarrow [0, 1]$$

defined by

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

More precisely,

$$F(x) = \mathbb{P}(\{\omega \in S : X(\omega) \leq x\}).$$

In fact, for any $B \subseteq \mathbb{R}$, we write $\{X \in B\}$ instead of $\{\omega \in S : X(\omega) \in B\}$.

Proposition 1 (Basic properties of a CDF). *Let F be the CDF of a random variable X . Then:*

(1)

$$0 \leq F(x) \leq 1 \quad \text{for every } x \in \mathbb{R}.$$

(2)

$$\lim_{x \rightarrow -\infty} F(x) = 0.$$

(3)

$$\lim_{x \rightarrow \infty} F(x) = 1.$$

(4) F is increasing. That is, if $x \leq y$, then

$$F(x) \leq F(y).$$

Proof. We prove each property separately.

(1) Proof that $F(x) \in [0, 1]$.

Fix $x \in \mathbb{R}$. By definition,

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

The set $\{X \leq x\}$ is an event in $\mathcal{F}$. Since $\mathbb{P}$ is a probability measure, every event has probability between 0 and 1. Therefore,

$$0 \leq \mathbb{P}(X \leq x) \leq 1.$$

Hence,

$$0 \leq F(x) \leq 1.$$

(2) Proof that $\lim_{x \rightarrow -\infty} F(x) = 0$.

For each $n \in \mathbb{N}$, define

$$A_n = \{X \leq -n\}.$$

As n increases, the number $-n$ becomes smaller. Therefore, the events become smaller, and thus, we have a decreasing chain of events.

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$$

Reason: If $\omega \in A_{n+1}$, then $X(\omega) \leq -(n+1) < -n$, so $X(\omega) \leq -n$. Thus, $\omega \in A_n$, and hence, $A_{n+1} \subseteq A_n$.

Now consider the intersection

$$\bigcap_{n=1}^{\infty} A_n = \bigcap_{n=1}^{\infty} \{X \leq -n\}.$$

This event means that X is less than or equal to $-n$ for every positive integer n . But no real number can be less than or equal to every number

$$-1, -2, -3, \dots$$

Since X is real-valued, this is impossible. Therefore,

$$\bigcap_{n=1}^{\infty} A_n = \emptyset.$$

By Theorem 2,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

Thus,

$$\mathbb{P}(\emptyset) = \lim_{n \rightarrow \infty} \mathbb{P}(X \leq -n).$$

Since $\mathbb{P}(\emptyset) = 0$, we get

$$\lim_{n \rightarrow \infty} F(-n) = 0.$$

This shows that along the sequence $x = -n$,

$$F(x) \rightarrow 0.$$

Since F is increasing, this implies the full limit

$$\lim_{x \rightarrow -\infty} F(x) = 0.$$

(3) Proof that $\lim_{x \rightarrow \infty} F(x) = 1$.

For each $n \in \mathbb{N}$, define

$$B_n = \{X \leq n\}.$$

As n increases, the number n becomes larger. Therefore, the events become larger:

$$B_1 \subseteq B_2 \subseteq B_3 \subseteq \cdots.$$

Reason: If $\omega \in B_n$, then $X(\omega) \leq n < n+1$, so $X(\omega) \leq n+1$. Thus, $\omega \in B_{n+1}$. Hence, $B_n \subseteq B_{n+1}$. Now consider the union.

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} \{X \leq n\}.$$

Since X is real-valued, for every outcome $\omega \in S$, the number $X(\omega)$ is a real number. Therefore, there exists some positive integer n such that

$$X(\omega) \leq n.$$

Hence every $\omega \in S$ belongs to at least one of the events B_n . Therefore,

$$\bigcup_{n=1}^{\infty} B_n = S.$$

By Theorem 1,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(B_n).$$

Thus,

$$\mathbb{P}(S) = \lim_{n \rightarrow \infty} \mathbb{P}(X \leq n).$$

Since $\mathbb{P}(S) = 1$, we get

$$\lim_{n \rightarrow \infty} F(n) = 1.$$

This shows that along the sequence $x = n$,

$$F(x) \rightarrow 1.$$

Since F is increasing and bounded above by 1, this implies the full limit

$$\lim_{x \rightarrow \infty} F(x) = 1.$$

(4) Proof that F is increasing.

Let $x, y \in \mathbb{R}$ and suppose

$$x \leq y.$$

We claim that

$$\{X \leq x\} \subseteq \{X \leq y\}.$$

Indeed, if $\omega \in \{X \leq x\}$, then

$$X(\omega) \leq x.$$

Since $x \leq y$, we have

$$X(\omega) \leq y.$$

Therefore,

$$\omega \in \{X \leq y\}.$$

So,

$$\{X \leq x\} \subseteq \{X \leq y\}.$$

By monotonicity of probability,

$$\mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y).$$

Therefore,

$$F(x) \leq F(y).$$

Hence F is increasing. □

2 Linear Algebra Background

The expectation of a random vector will itself be a vector, while its variances and covariances will be organized into a matrix called the *covariance matrix*.

We first review the basic linear algebra notation that we will need.

Vectors

A **vector** in $\mathbb{R}^n$ is an ordered collection of n real numbers. We usually write vectors as column vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

For example,

$$\mathbf{x} = \begin{pmatrix} 2 \\ -1 \\ 4 \end{pmatrix} \in \mathbb{R}^3.$$

Two vectors are equal if and only if all of their corresponding components are equal:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \iff x_i = y_i \text{ for every } i.$$

Vectors can be added componentwise:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}.$$

Similarly, multiplication by a scalar $c \in \mathbb{R}$ is defined by

$$c \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} cx_1 \\ \vdots \\ cx_n \end{pmatrix}.$$

Matrices

A **matrix** is a rectangular array of numbers. For example,

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

has 2 rows and 3 columns, so we say that A is a

$$2 \times 3$$

matrix, or belongs to $\mathbb{R}^{2 \times 3}$.

More generally, an $m \times n$ matrix has the form

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

The notation a_{ij} denotes the entry in row i and column j .

Thus, if

$$A = \begin{pmatrix} 2 & 5 & 7 \\ 1 & 3 & 4 \end{pmatrix},$$

then

$$a_{11} = 2, \quad a_{12} = 5, \quad a_{23} = 4.$$

A matrix with the same number of rows and columns is called a **square matrix**.

For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

is a 2×2 square matrix.

Matrix Addition and Scalar Multiplication

Two matrices can be added only when they have the same dimensions. Addition is performed entry by entry.

For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}.$$

Similarly,

$$cA$$

means that every entry of A is multiplied by c . For example,

$$3 \begin{pmatrix} 1 & -2 \\ 4 & 5 \end{pmatrix} = \begin{pmatrix} 3 & -6 \\ 12 & 15 \end{pmatrix}.$$

Transpose of a Matrix

Definition 3 (Transpose). If A is an $m \times n$ matrix, its **transpose**, denoted by A^T , is the $n \times m$ matrix obtained by turning the rows of A into columns.

Equivalently,

$$(A^T)_{ij} = A_{ji}.$$

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

then

$$A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

In particular, if

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

then

$$\mathbf{x}^T = (x_1 \quad x_2 \quad x_3).$$

Thus, transposing a column vector turns it into a row vector.

Note: Vectors are special cases of matrices, i.e., $\curvearrowright \in \mathbb{R}^n = \mathbb{R}^{n \times 1}$, meaning that a vector with n entries is a matrix with n columns and 1 row.

Multiplication of Vectors

There are two ways to multiply $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, namely, the inner product (defined by $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y}$) and the outer product (define by $\mathbf{x} \times \mathbf{y} = \mathbf{x} \mathbf{y}^T$).

Suppose

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Since $\mathbf{x} \in \mathbb{R}^n$ is an $n \times 1$ column vector, its transpose $\mathbf{x}^T$ is a $1 \times n$ row vector. Therefore, $\mathbf{x}^T$ can be multiplied by the $n \times 1$ vector $\mathbf{y}$, and the result has dimension

$$(1 \times n)(n \times 1) = (1 \times 1),$$

which is a scalar.

Thus,

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i.$$

This is called the **dot product** or **inner product**.

For example,

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}^T \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} = 1(4) + 2(5) + 3(6) = 32.$$

Notice, however, that reversing the multiplication produces something completely different:

$$\mathbf{xy}^T.$$

This is an $n \times n$ matrix:

$$\mathbf{xy}^T = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_ny_1 & x_ny_2 & \cdots & x_ny_n \end{pmatrix}.$$

This is called the **outer product**.

For example,

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 6 & 8 \end{pmatrix}.$$

This distinction will be especially important when we define the covariance matrix of a random vector.

Matrix Multiplication

Matrix multiplication is not performed entry by entry.

Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix. Then the product

$$AB$$

is defined and is an $m \times p$ matrix.

Matrix multiplication is not always performable. It can only be done when the number of columns of A is equal to the number of rows of B . Schematically,

$$(m \times n)(n \times p) \longrightarrow (m \times p).$$

The two *inside dimensions* must agree.

For example,

$$(2 \times 3)(3 \times 4) \longrightarrow (2 \times 4).$$

However,

$$(2 \times 3)(2 \times 4)$$

is not defined because the inside dimensions 3 and 2 do not agree.

To compute the $(i, j)^{th}$ entry of AB (the cell in i^{th} row and j^{th} column), multiply (inner product) the i^{th} row of A by the j^{th} column of B .

For example,

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}.$$

Then

$$AB = \begin{pmatrix} 1(5) + 2(7) & 1(6) + 2(8) \\ 3(5) + 4(7) & 3(6) + 4(8) \end{pmatrix},$$

so

$$AB = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}.$$

In general, if

$$A = (a_{ij})_{m \times n}, \quad B = (b_{ij})_{n \times p},$$

and

$$C = AB,$$

then

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}.$$

This is exactly the dot product of row i of A with column j of B .

Matrix–Vector Multiplication

A particularly important case occurs when a matrix is multiplied by a vector.

If

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

then

$$A\mathbf{x} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}.$$

More generally, if A is $m \times n$ and $\mathbf{x} \in \mathbb{R}^n$, then

$$A\mathbf{x} \in \mathbb{R}^m.$$

Some Important Rules for Matrix Multiplication

Whenever the dimensions are appropriate,

$$A(B + C) = AB + AC,$$

and

$$(A + B)C = AC + BC.$$

Matrix multiplication is also associative:

$$A(BC) = (AB)C.$$

However, matrix multiplication is generally **not commutative**. That is,

$$AB \neq BA$$

in general.

In fact, it is possible for AB to be defined while BA is not defined.

For transposes,

$$(A^T)^T = A$$

and

$$(AB)^T = B^T A^T.$$

Notice that the order is reversed when taking the transpose of a product.

Determinants

The **determinant** is a number associated with a square matrix. For a square matrix A , we denote its determinant by

$$\det(A) \quad \text{or} \quad |A|.$$

We begin with the 2×2 case.

Determinant of a 2×2 Matrix

If

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then

$$\det(A) = ad - bc.$$

For example,

$$A = \begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}.$$

Then

$$\det(A) = 2(4) - 3(1) = 8 - 3 = 5.$$

Determinant of a 3×3 Matrix

Consider

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}.$$

There are multiple ways to compute the determinant of a 3×3 matrix. One way is to expand along the first row, i.e.,

$$\det(A) = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}.$$

Using the 2×2 determinant formula,

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg).$$

This method gives a recursive rule to compute the determinant of any $n \times n$ matrix to a 2×2 matrix (by expanding on the first row with alternating signs).

Notice the alternating signs

$$+, \quad -, \quad +.$$

For example, let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 1 & 0 & 6 \end{pmatrix}.$$

Expanding along the first row gives

$$\begin{aligned} \det(A) &= 1 \begin{vmatrix} 4 & 5 \\ 0 & 6 \end{vmatrix} - 2 \begin{vmatrix} 0 & 5 \\ 1 & 6 \end{vmatrix} + 3 \begin{vmatrix} 0 & 4 \\ 1 & 0 \end{vmatrix} \\ &= 1(24) - 2(0 - 5) + 3(0 - 4) \\ &= 24 + 10 - 12 \\ &= 22. \end{aligned}$$

Minors and Cofactors

Let A be an $n \times n$ matrix.

For an entry a_{ij} , delete row i and column j . The resulting $(n-1) \times (n-1)$ matrix is denoted by

$$A_{ij}.$$

The number

$$M_{ij} = \det(A_{ij})$$

is called the **minor** corresponding to a_{ij} .

The **cofactor** of a_{ij} is

$$C_{ij} = (-1)^{i+j} M_{ij}.$$

The signs therefore follow the checkerboard pattern

$$\begin{pmatrix} + & - & + & - \cdots \\ - & + & - & + \cdots \\ + & - & + & - \cdots \\ - & + & - & + \cdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Determinant of an $n \times n$ Matrix

For an $n \times n$ matrix

$$A = (a_{ij}),$$

the determinant can be calculated by expanding along any row or any column.

For example, expanding along row i gives

$$\det(A) = \sum_{j=1}^n a_{ij} C_{ij} = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}).$$

Thus, computing an $n \times n$ determinant reduces to computing $(n-1) \times (n-1)$ determinants, which reduce to $(n-2) \times (n-2)$ determinants, and so on, until we reach the 2×2 formula.

Similarly, expanding along column j gives

$$\det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}).$$

Remark 3. When computing a determinant by cofactor expansion, it is usually convenient to choose a row or column containing many zeros.

3 Random Vectors, Expectations, and Covariance Matrices

We now use the preceding notation to describe several random variables simultaneously.

Random Vectors

Definition 4 (Random vector). Let

$$X_1, X_2, \dots, X_n$$

be random variables defined on the same probability space. The column vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

is called an n -dimensional **random vector**.

For example,

$$\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

is a two-dimensional random vector.

For each outcome $\omega \in S$,

$$\mathbf{X}(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_n(\omega) \end{pmatrix} \in \mathbb{R}^n.$$

Thus, instead of producing one random number, a random vector produces a vector of random numbers.

Expected Value of a Random Vector

Recall that the expectation of a single random variable X is a number $E(X)$.

For a random vector, we simply take the expectation of each component.

Definition 5 (Expected value of a random vector). Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

where each X_i has a finite expectation. Then

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

We often denote this vector by

$$\boldsymbol{\mu} = E(\mathbf{X}).$$

Thus,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \mu_i = E(X_i).$$

For example, if

$$E(X) = 2, \quad E(Y) = -1,$$

then

$$E \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Linearity of Expectation in Matrix Form

Suppose A is a fixed $m \times n$ matrix and $\mathbf{X} \in \mathbb{R}^n$ is a random vector. Then

$$A\mathbf{X}$$

is another random vector.

By linearity of expectation,

$$E(A\mathbf{X}) = A E(\mathbf{X}).$$

More generally, for a fixed vector $\mathbf{b}$,

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

From Variance to the Covariance Matrix

For a single random variable X , recall that

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

We would like to generalize this definition to a random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

Let

$$\boldsymbol{\mu} = E(\mathbf{X}).$$

Then the centered random vector is

$$\mathbf{X} - \boldsymbol{\mu} = \begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_n - \mu_n \end{pmatrix}.$$

A tempting expression would be

$$(\mathbf{X} - \boldsymbol{\mu})^T (\mathbf{X} - \boldsymbol{\mu}),$$

but notice its dimensions:

$$(1 \times n)(n \times 1) = 1 \times 1.$$

Therefore, this produces only one number:

$$\sum_{i=1}^n (X_i - \mu_i)^2.$$

Instead, reverse the order:

$$(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T.$$

Now the dimensions are

$$(n \times 1)(1 \times n) = n \times n,$$

so the result is a matrix containing every pairwise product:

$$(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T = \begin{pmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_n - \mu_n) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & (X_n - \mu_n)(X_2 - \mu_2) & \cdots & (X_n - \mu_n)^2 \end{pmatrix}.$$

Taking expectations entry by entry gives exactly the variances and covariances.

Covariance Matrix

Definition 6 (Covariance matrix). Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

be a random vector with finite second moments, and let

$$\boldsymbol{\mu} = E(\mathbf{X}).$$

The **covariance matrix** of $\mathbf{X}$ is

$$Cov(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T].$$

It is also commonly denoted by

$$\Sigma.$$

The (i, j) entry of this matrix is

$$E[(X_i - \mu_i)(X_j - \mu_j)].$$

But this is exactly

$$Cov(X_i, X_j).$$

Therefore,

$$\Sigma = Cov(\mathbf{X}) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Var(X_n) \end{pmatrix}.$$

Thus:

- The **diagonal entries** are the variances:

$$\Sigma_{ii} = \text{Var}(X_i).$$

- The **off-diagonal entries** are the covariances:

$$\Sigma_{ij} = \text{Cov}(X_i, X_j), \quad i \neq j.$$

The Two-Dimensional Case

For

$$\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad E(\mathbf{X}) = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},$$

we have

$$\mathbf{X} - E(\mathbf{X}) = \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}.$$

Therefore,

$$\begin{aligned} & (\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T \\ &= \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix} \begin{pmatrix} X - \mu_X & Y - \mu_Y \end{pmatrix} \\ &= \begin{pmatrix} (X - \mu_X)^2 & (X - \mu_X)(Y - \mu_Y) \\ (Y - \mu_Y)(X - \mu_X) & (Y - \mu_Y)^2 \end{pmatrix}. \end{aligned}$$

Taking expectations,

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{pmatrix}.$$

Since

$$\text{Cov}(X, Y) = \text{Cov}(Y, X),$$

this becomes

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{pmatrix}.$$

An Equivalent Formula for the Covariance Matrix

Recall the scalar identity

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

There is an analogous matrix identity.

Proposition 2. *Let*

$$\boldsymbol{\mu} = E(\mathbf{X}).$$

Then

$$\text{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

Equivalently,

$$\text{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^T) - E(\mathbf{X})E(\mathbf{X})^T.$$

Proof. Starting from the definition,

$$\begin{aligned} \text{Cov}(\mathbf{X}) &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \\ &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X}^T - \boldsymbol{\mu}^T)] \\ &= E[\mathbf{X}\mathbf{X}^T - \mathbf{X}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{X}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T]. \end{aligned}$$

Since $\boldsymbol{\mu}$ is a fixed vector,

$$\text{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^T) - E(\mathbf{X})\boldsymbol{\mu}^T - \boldsymbol{\mu}E(\mathbf{X})^T + \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

But

$$E(\mathbf{X}) = \boldsymbol{\mu},$$

so

$$\begin{aligned} \text{Cov}(\mathbf{X}) &= E(\mathbf{X}\mathbf{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T \\ &= E(\mathbf{X}\mathbf{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T. \end{aligned}$$

□

Symmetry of the Covariance Matrix

Proposition 3. *The covariance matrix is symmetric:*

$$\text{Cov}(\mathbf{X})^T = \text{Cov}(\mathbf{X}).$$

Proof. The (i, j) entry of $\text{Cov}(\mathbf{X})$ is

$$\text{Cov}(X_i, X_j).$$

But

$$\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i).$$

Therefore,

$$\Sigma_{ij} = \Sigma_{ji},$$

which means that Σ is symmetric. □

Covariance Matrix Under Linear Transformations

The matrix notation becomes especially useful when we transform a random vector.

Proposition 4. *Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector with covariance matrix*

$$\Sigma = \text{Cov}(\mathbf{X}),$$

let A be a fixed $m \times n$ matrix, and let $\mathbf{b} \in \mathbb{R}^m$ be fixed. Then

$$\text{Cov}(A\mathbf{X} + \mathbf{b}) = A\Sigma A^T.$$

Proof. Let

$$\mathbf{Y} = A\mathbf{X} + \mathbf{b}.$$

Then

$$E(\mathbf{Y}) = AE(\mathbf{X}) + \mathbf{b}.$$

Hence,

$$\begin{aligned}\mathbf{Y} - E(\mathbf{Y}) &= A\mathbf{X} + \mathbf{b} - AE(\mathbf{X}) - \mathbf{b} \\ &= A(\mathbf{X} - E(\mathbf{X})).\end{aligned}$$

Therefore,

$$\begin{aligned}Cov(\mathbf{Y}) &= E[(\mathbf{Y} - E(\mathbf{Y}))(\mathbf{Y} - E(\mathbf{Y}))^T] \\ &= E[A(\mathbf{X} - E(\mathbf{X}))(A(\mathbf{X} - E(\mathbf{X})))^T].\end{aligned}$$

Using

$$(AB)^T = B^T A^T,$$

we obtain

$$\begin{aligned}Cov(\mathbf{Y}) &= E[A(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T A^T] \\ &= AE[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T] A^T \\ &= A\Sigma A^T.\end{aligned}$$

□

Variance of a Linear Combination

An important special case is a linear combination

$$Y = a_1 X_1 + \cdots + a_n X_n.$$

Define

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}.$$

Then

$$Y = \mathbf{a}^T \mathbf{X}.$$

Since $\mathbf{a}^T$ is a $1 \times n$ matrix, the preceding proposition gives

$$Var(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \Sigma \mathbf{a}.$$

Expanding this expression,

$$Var\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 Var(X_i) + 2 \sum_{i < j} a_i a_j Cov(X_i, X_j).$$

For two random variables,

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab Cov(X, Y).$$

Thus, the familiar variance formula for a linear combination is contained naturally inside the covariance-matrix notation.

4 The Central Limit Theorem

The Central Limit Theorem is one of the most important results in probability. It explains why the normal distribution appears so frequently when we study sums and averages of many independent random variables.

The interesting part of it is that the individual random variables do *not* need to be normally distributed. Under suitable assumptions, their properly standardized sum becomes approximately normal as the number of terms becomes large.

Convergence in Distribution

Before stating the Central Limit Theorem, we recall the relevant notion of convergence.

Definition 7 (Convergence in distribution). Let $X_1, X_2, \dots$ and X be random variables with CDFs $F_{X_1}, F_{X_2}, \dots$ and F_X , respectively.

We say that

$$X_n \xrightarrow{d} X$$

if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

for every $x \in \mathbb{R}$ at which F_X is continuous.

In particular, since the CDF of a normal random variable is continuous everywhere,

$$X_n \xrightarrow{d} N(0, 1)$$

means that

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n \leq x) = \Phi(x) \quad \text{for every } x \in \mathbb{R},$$

where Φ is the CDF of the standard normal distribution.

The Lindeberg–Lévy Central Limit Theorem

Theorem 3 (Central Limit Theorem). *Let*

$$X_1, X_2, \dots$$

be independent and identically distributed random variables satisfying

$$E(X_i) = \mu$$

and

$$\text{Var}(X_i) = \sigma^2, \quad 0 < \sigma^2 < \infty.$$

Define

$$S_n = X_1 + \dots + X_n.$$

Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1).$$

Equivalently, if

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i,$$

then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

That is, for every $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x \right) = \Phi(x).$$

Remark 4 (Assumptions of the theorem). The theorem requires the following assumptions.

- (1) The random variables

$$X_1, X_2, \dots$$

are independent.

- (2) They are identically distributed.

- (3) Their common mean exists and is finite:

$$E(X_i) = \mu \in \mathbb{R}.$$

- (4) Their common variance exists, is finite, and is positive:

$$0 < \text{Var}(X_i) = \sigma^2 < \infty.$$

No assumption is made that the X_i themselves are normally distributed.

Characteristic Functions

We prove the Central Limit Theorem using characteristic functions.

Definition 8 (Characteristic function). The **characteristic function** of a random variable X is

$$\varphi_X(t) = E(e^{itX}), \quad t \in \mathbb{R},$$

where

$$i = \sqrt{-1}.$$

Since

$$|e^{itX}| = 1,$$

the expectation

$$E(e^{itX})$$

always exists. Thus, unlike an MGF, a characteristic function exists for every random variable.

Using Euler's formula,

$$e^{itX} = \cos(tX) + i \sin(tX),$$

we may also write

$$\varphi_X(t) = E[\cos(tX)] + iE[\sin(tX)].$$

Characteristic Functions and Independence

Proposition 5. *If X and Y are independent, then*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

Proof. By definition,

$$\begin{aligned}\varphi_{X+Y}(t) &= E\left(e^{it(X+Y)}\right) \\ &= E\left(e^{itX}e^{itY}\right).\end{aligned}$$

Since X and Y are independent, the random variables

$$e^{itX} \quad \text{and} \quad e^{itY}$$

are also independent. Therefore,

$$\begin{aligned}E\left(e^{itX}e^{itY}\right) &= E(e^{itX})E(e^{itY}) \\ &= \varphi_X(t)\varphi_Y(t).\end{aligned}$$

Hence,

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

□

More generally, if

$$X_1, \dots, X_n$$

are independent, then

$$\varphi_{X_1+\dots+X_n}(t) = \prod_{j=1}^n \varphi_{X_j}(t).$$

If, in addition, the random variables are identically distributed, then

$$\varphi_{X_1+\dots+X_n}(t) = [\varphi_{X_1}(t)]^n.$$

Characteristic Functions Under Linear Transformations

Proposition 6. *If*

$$Y = aX + b,$$

where $a, b \in \mathbb{R}$, then

$$\varphi_Y(t) = e^{itb}\varphi_X(at).$$

Proof. By definition,

$$\begin{aligned}\varphi_Y(t) &= E(e^{itY}) \\ &= E(e^{it(aX+b)}) \\ &= E(e^{itb}e^{iatX}) \\ &= e^{itb}E(e^{i(at)X}) \\ &= e^{itb}\varphi_X(at).\end{aligned}$$

□

A Second-Order Expansion of a Characteristic Function

The following proposition is the main ingredient in the proof of the Central Limit Theorem.

Proposition 7. *Suppose Y satisfies*

$$E(Y) = 0$$

and

$$E(Y^2) = 1.$$

Then, as $t \rightarrow 0$,

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2).$$

Proof. We begin with the second-order Taylor expansion of the complex exponential:

$$e^{ix} = 1 + ix - \frac{x^2}{2} + o(x^2) \quad \text{as } x \rightarrow 0.$$

To use this expansion inside an expectation, we make the remainder precise. Define

$$g(x) = \begin{cases} \frac{e^{ix} - 1 - ix + \frac{x^2}{2}}{x^2}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Since

$$e^{ix} = 1 + ix - \frac{x^2}{2} + o(x^2),$$

we have

$$\lim_{x \rightarrow 0} g(x) = 0.$$

Therefore,

$$e^{ix} = 1 + ix - \frac{x^2}{2} + x^2 g(x),$$

where

$$g(x) \rightarrow 0 \quad \text{as } x \rightarrow 0.$$

We also need the fact that g is bounded.

For $|x| \geq 1$,

$$\begin{aligned} |g(x)| &\leq \frac{|e^{ix}| + 1 + |x| + \frac{x^2}{2}}{x^2} \\ &= \frac{2 + |x| + \frac{x^2}{2}}{x^2} \\ &\leq 2 + 1 + \frac{1}{2} \\ &= \frac{7}{2}. \end{aligned}$$

Near zero, g is continuous and converges to zero, so it is also bounded on a neighborhood of zero. Hence, there exists some constant $C < \infty$ such that

$$|g(x)| \leq C \quad \text{for every } x \in \mathbb{R}.$$

Now substitute

$$x = tY.$$

We obtain

$$e^{itY} = 1 + itY - \frac{t^2 Y^2}{2} + t^2 Y^2 g(tY).$$

Taking expectations gives

$$\begin{aligned}\varphi_Y(t) &= E(e^{itY}) \\ &= 1 + itE(Y) - \frac{t^2}{2}E(Y^2) + t^2E[Y^2g(tY)].\end{aligned}$$

Since

$$E(Y) = 0$$

and

$$E(Y^2) = 1,$$

this becomes

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + t^2E[Y^2g(tY)].$$

We now show that

$$E[Y^2g(tY)] \rightarrow 0 \quad \text{as } t \rightarrow 0.$$

For every fixed value of Y ,

$$tY \rightarrow 0,$$

and therefore

$$g(tY) \rightarrow 0.$$

Hence,

$$Y^2g(tY) \rightarrow 0$$

pointwise.

Furthermore, since g is bounded,

$$|Y^2g(tY)| \leq CY^2.$$

But

$$E(Y^2) = 1 < \infty.$$

Therefore, by the Dominated Convergence Theorem,

$$E[Y^2g(tY)] \rightarrow 0.$$

Thus,

$$E[Y^2g(tY)] = o(1),$$

and consequently,

$$t^2E[Y^2g(tY)] = o(t^2).$$

We conclude that

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2).$$

□

Characteristic Function of the Standard Normal

We will also need the characteristic function of a standard normal random variable.

Proposition 8. *If*

$$Z \sim N(0, 1),$$

then

$$\varphi_Z(t) = e^{-t^2/2}.$$

Proof. Since

$$Z \sim N(0, 1),$$

its density is

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

Define

$$\varphi_Z(t) = E(e^{itZ}).$$

We differentiate with respect to t :

$$\varphi'_Z(t) = E(iZe^{itZ}).$$

Therefore,

$$\varphi'_Z(t) = \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{itz} e^{-z^2/2} dz.$$

Notice that

$$\frac{d}{dz} e^{-z^2/2} = -ze^{-z^2/2}.$$

Thus,

$$ze^{-z^2/2} = -\frac{d}{dz} e^{-z^2/2}.$$

Hence,

$$\varphi'_Z(t) = -\frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itz} \frac{d}{dz} (e^{-z^2/2}) dz.$$

Using integration by parts,

$$\begin{aligned} \varphi'_Z(t) &= -\frac{i}{\sqrt{2\pi}} \left[e^{itz} e^{-z^2/2} \right]_{-\infty}^{\infty} \\ &\quad + \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ite^{itz} e^{-z^2/2} dz. \end{aligned}$$

Since

$$e^{-z^2/2} \rightarrow 0 \quad \text{as } |z| \rightarrow \infty,$$

the boundary term is zero. Therefore,

$$\begin{aligned} \varphi'_Z(t) &= \frac{i^2 t}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itz} e^{-z^2/2} dz \\ &= -\frac{t}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itz} e^{-z^2/2} dz \\ &= -t\varphi_Z(t). \end{aligned}$$

Thus, φ_Z satisfies the differential equation

$$\varphi'_Z(t) = -t\varphi_Z(t).$$

Equivalently,

$$\frac{\varphi'_Z(t)}{\varphi_Z(t)} = -t.$$

Integrating,

$$\log \varphi_Z(t) = -\frac{t^2}{2} + C.$$

Since every characteristic function satisfies

$$\varphi_Z(0) = 1,$$

we have

$$0 = C.$$

Therefore,

$$\varphi_Z(t) = e^{-t^2/2}.$$

□

The Characteristic-Function Convergence Theorem

We use the following standard result.

Theorem 4 (Lévy's Continuity Theorem). *Let $Y_1, Y_2, \dots$ be random variables with characteristic functions*

$$\varphi_{Y_1}, \varphi_{Y_2}, \dots$$

Suppose there exists a random variable Y with characteristic function φ_Y such that

$$\varphi_{Y_n}(t) \longrightarrow \varphi_Y(t)$$

for every $t \in \mathbb{R}$.

Then

$$Y_n \xrightarrow{d} Y.$$

Thus, to prove convergence in distribution, it is enough to prove pointwise convergence of the corresponding characteristic functions.

Proof of the Central Limit Theorem

Proof of Theorem 3. Let

$$X_1, X_2, \dots$$

be independent and identically distributed with

$$E(X_i) = \mu$$

and

$$\text{Var}(X_i) = \sigma^2, \quad 0 < \sigma^2 < \infty.$$

We want to prove that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1),$$

where

$$S_n = X_1 + \cdots + X_n.$$

Step 1: Standardize the individual random variables.

Define

$$Y_i = \frac{X_i - \mu}{\sigma}.$$

Because the X_i are independent and identically distributed, the Y_i are also independent and identically distributed.

Moreover,

$$\begin{aligned} E(Y_i) &= E\left(\frac{X_i - \mu}{\sigma}\right) \\ &= \frac{1}{\sigma} (E(X_i) - \mu) \\ &= 0. \end{aligned}$$

Also,

$$\begin{aligned} \text{Var}(Y_i) &= \text{Var}\left(\frac{X_i - \mu}{\sigma}\right) \\ &= \frac{1}{\sigma^2} \text{Var}(X_i) \\ &= 1. \end{aligned}$$

Since

$$E(Y_i) = 0,$$

we therefore have

$$E(Y_i^2) = \text{Var}(Y_i) = 1.$$

Now observe that

$$\begin{aligned} \frac{S_n - n\mu}{\sigma\sqrt{n}} &= \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \\ &= \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i. \end{aligned}$$

Define

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i.$$

Thus, it is enough to prove that

$$Z_n \xrightarrow{d} N(0, 1).$$

Step 2: Find the characteristic function of Z_n .

Let

$$\varphi_Y(t) = E(e^{itY_1})$$

denote the common characteristic function of the Y_i .

Then

$$\begin{aligned}\varphi_{Z_n}(t) &= E \left[\exp \left(it \frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \right) \right] \\ &= E \left[\exp \left(i \frac{t}{\sqrt{n}} Y_1 \right) \cdots \exp \left(i \frac{t}{\sqrt{n}} Y_n \right) \right].\end{aligned}$$

Because

$$Y_1, \dots, Y_n$$

are independent,

$$\varphi_{Z_n}(t) = \prod_{j=1}^n E \left[\exp \left(i \frac{t}{\sqrt{n}} Y_j \right) \right].$$

Because the Y_j are identically distributed,

$$\varphi_{Z_n}(t) = \left[\varphi_Y \left(\frac{t}{\sqrt{n}} \right) \right]^n.$$

Step 3: Apply the second-order expansion.

From Proposition 7,

$$\varphi_Y(u) = 1 - \frac{u^2}{2} + o(u^2) \quad \text{as } u \rightarrow 0.$$

Set

$$u = \frac{t}{\sqrt{n}}.$$

For each fixed t ,

$$\frac{t}{\sqrt{n}} \rightarrow 0.$$

Therefore,

$$\begin{aligned}\varphi_Y \left(\frac{t}{\sqrt{n}} \right) &= 1 - \frac{1}{2} \left(\frac{t}{\sqrt{n}} \right)^2 + o \left(\left(\frac{t}{\sqrt{n}} \right)^2 \right) \\ &= 1 - \frac{t^2}{2n} + o \left(\frac{1}{n} \right).\end{aligned}$$

Hence,

$$\varphi_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o \left(\frac{1}{n} \right) \right]^n.$$

We now evaluate this limit.

Write

$$a_n = -\frac{t^2}{2n} + o\left(\frac{1}{n}\right).$$

Then

$$a_n \rightarrow 0$$

and

$$na_n \rightarrow -\frac{t^2}{2}.$$

Also,

$$a_n = O\left(\frac{1}{n}\right),$$

so

$$na_n^2 = O\left(\frac{1}{n}\right) \rightarrow 0.$$

Using the expansion

$$\log(1+x) = x + o(x) \quad \text{as } x \rightarrow 0,$$

or, more precisely,

$$\log(1+x) = x + O(x^2),$$

we obtain

$$\begin{aligned} \log[(1+a_n)^n] &= n \log(1+a_n) \\ &= n[a_n + O(a_n^2)] \\ &= na_n + O(na_n^2). \end{aligned}$$

Taking the limit,

$$na_n \rightarrow -\frac{t^2}{2}$$

and

$$na_n^2 \rightarrow 0.$$

Therefore,

$$\log[(1+a_n)^n] \rightarrow -\frac{t^2}{2}.$$

Exponentiating,

$$(1+a_n)^n \rightarrow e^{-t^2/2}.$$

Consequently,

$$\lim_{n \rightarrow \infty} \varphi_{Z_n}(t) = e^{-t^2/2}.$$

Step 4: Identify the limiting distribution.

We proved above that if

$$Z \sim N(0, 1),$$

then

$$\varphi_Z(t) = e^{-t^2/2}.$$

Thus,

$$\varphi_{Z_n}(t) \longrightarrow \varphi_Z(t) \quad \text{for every } t \in \mathbb{R}.$$

By Lévy's Continuity Theorem,

$$Z_n \xrightarrow{d} Z.$$

Since

$$Z \sim N(0, 1),$$

we obtain

$$Z_n \xrightarrow{d} N(0, 1).$$

Finally,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}},$$

so

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1).$$

This proves the result. □

Equivalent Form in Terms of the Sample Mean

Since

$$\bar{X}_n = \frac{S_n}{n},$$

we have

$$\begin{aligned} \frac{S_n - n\mu}{\sigma\sqrt{n}} &= \frac{n\bar{X}_n - n\mu}{\sigma\sqrt{n}} \\ &= \frac{n(\bar{X}_n - \mu)}{\sigma\sqrt{n}} \\ &= \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}. \end{aligned}$$

Therefore, the Central Limit Theorem may equivalently be written as

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

Another equivalent way to express the approximation is

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

for large n .

Similarly,

$$S_n \approx N(n\mu, n\sigma^2)$$

for large n .