Statistics

Personal Study Notes and Formal Derivations

Victor Alves

2026-04

Contents

Disclaimer

- **Incomplete Document:** This file constitutes supporting material under active development. Several sections and asymptotic derivations are still being revised, expanded, and supplemented.
- **AI-Assisted Construction:** This material was structured, reviewed, and expanded with the assistance of Artificial Intelligence models for didactic organization and Markdown/LaTeX formatting rigor.
- **Margin of Error:** Due to the technical nature of the matrix and asymptotic proofs, the text may contain typographical errors, algebraic omissions, or theoretical inaccuracies not yet reviewed by the author. It should not be used as the sole definitive bibliographic source.

Chapter 1: Foundations of Probability

1.1 Sample Space and Set Operations

1.1.1 Motivation

The central problem addressed in this chapter is the **formalization of uncertainty**. Without a rigorous mathematical framework for describing random phenomena—such as financial market fluctuations, epidemiological outcomes, or the lifetime of engineered components—analysis would remain confined to

subjective intuition. The language of probability provides the necessary structure to apply tools from measure theory, calculus, and linear algebra, enabling consistent prediction and inference.

Geometric Interpretation in $\mathbb{R}^2$: Visualize the sample space Ω as a bounded region in the plane, say the unit square $[0,1]^2$. Each point $\omega \in \Omega$ corresponds to a specific outcome of the random experiment. An *event* A is a measurable subregion of this square. Under a uniform probability distribution, $\mathbb{P}(A)$ is simply the **ratio of the area of A to the total area of Ω**. In this framework, logical operations such as "A or B" correspond to the *union* of regions, while "A and B" correspond to their *intersection*.

1.1.2 Basic Definitions

Let $\mathcal{E}$ denote a random experiment. We define the following fundamental objects:

1. **Sample Space (Ω):** A non-empty set containing all possible outcomes ω of $\mathcal{E}$.
2. **Event:** Any subset $A \subseteq \Omega$. We say that event A *occurs* if the realized outcome ω belongs to A.
3. **Containment:** For events $A, B \subseteq \Omega$,
$$A \subset B \iff \forall \omega \in A,\ \omega \in B.$$
4. **Equality:** Two events are equal if they contain precisely the same outcomes:
$$A = B \iff A \subset B \text{ and } B \subset A.$$

1.1.3 Fundamental Set Operations

For arbitrary events $A, B \subseteq \Omega$, we define:

1. **Union ($A \cup B$):** The event that at least one of A or B occurs:
$$A \cup B \equiv \{\omega \in \Omega \mid \omega \in A \text{ or } \omega \in B\}.$$
2. **Intersection ($A \cap B$):** The event that both A and B occur simultaneously:
$$A \cap B \equiv \{\omega \in \Omega \mid \omega \in A \text{ and } \omega \in B\}.$$
3. **Complement (A^c):** The event that A does **not** occur:
$$A^c \equiv \{\omega \in \Omega \mid \omega \notin A\}.$$
4. **Set Difference ($A \setminus B$):** The event that A occurs but B does not:
$$A \setminus B \equiv A \cap B^c.$$

1.1.4 Algebraic Properties of Set Operations

For any events $A, B, C \subseteq \Omega$, the following laws hold:

a. **Commutativity:**
$$A \cup B = B \cup A, \qquad A \cap B = B \cap A.$$

b. **Associativity:**
$$A \cup (B \cup C) = (A \cup B) \cup C, \qquad A \cap (B \cap C) = (A \cap B) \cap C.$$

c. **Distributive Laws:**

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

d. **De Morgan's Laws:**

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

1.1.5 Countable Operations and Partitions

The union and intersection operations extend naturally to countable collections of events. Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of subsets of Ω. Then:

$$\bigcup_{n=1}^{\infty} A_n = \{\omega \in \Omega \mid \exists n \in \mathbb{N} : \omega \in A_n\},$$
$$\bigcap_{n=1}^{\infty} A_n = \{\omega \in \Omega \mid \forall n \in \mathbb{N} : \omega \in A_n\}.$$

Example: Let $\Omega = (0, 1]$ and define $A_n = [\frac{1}{n}, 1]$. Then:

$$\bigcup_{n=1}^{\infty} A_n = (0, 1], \qquad \bigcap_{n=1}^{\infty} A_n = \{1\}.$$

Definition (Disjointness): Two events A and B are **disjoint** (or *mutually exclusive*) if $A \cap B = \emptyset$. A sequence $\{A_n\}_{n=1}^{\infty}$ is **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition (Partition): A collection $\{A_n\}_{n=1}^{\infty}$ is a **partition** of Ω if: 1. **Pairwise disjoint:** $A_i \cap A_j = \emptyset$ for all $i \neq j$; 2. **Exhaustive:** $\bigcup_{n=1}^{\infty} A_n = \Omega$.

1.2 σ-Algebras and Probability Measures

1.2.1 Motivation

The σ-algebra addresses a fundamental technical challenge: **delimiting which subsets of Ω are measurable**. In uncountably infinite sample spaces—such as $\mathbb{R}$—attempting to assign probabilities to *every* subset of Ω (i.e., using the power set 2^{Ω}) leads to logical inconsistencies, notably the failure of countable additivity due to the existence of non-measurable sets (a consequence of the Axiom of Choice).

Conceptually, a σ-algebra functions as an **information filter** or **resolution scale**. Returning to the geometric analogy: if Ω is a plane, the σ-algebra specifies which geometric figures possess a well-defined area. It guarantees closure under the operations we naturally require: if we can measure the area of two sets, we can also measure the area of their union, intersection, and complements.

1.2.2 Definition of a σ-Algebra

A collection $\mathcal{F}$ of subsets of Ω is called a **σ-algebra** (or **σ-field**) over Ω if it satisfies:

1. **Total set included:** $\Omega \in \mathcal{F}$.
2. **Closure under complementation:**

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

3. **Closure under countable unions:** For any sequence $\{A_n\}_{n=1}^{\infty}$ with $A_n \in \mathcal{F}$ for all n,

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

The pair $(\Omega, \mathcal{F})$ is called a **measurable space**.

Remark: Given Ω: - The **largest** σ-algebra is the power set 2^{Ω}. - The **smallest** σ-algebra is the trivial σ-field $\{\emptyset, \Omega\}$.

1.2.3 The Borel σ-Algebra

Let $\mathbb{R}^d$ denote d-dimensional Euclidean space.

- The **Borel σ-algebra** $\mathcal{B}(\mathbb{R}^d)$ is defined as the smallest σ-algebra containing all open subsets of $\mathbb{R}^d$.
- For $d = 1$, $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open intervals (a, b) with $a < b$.

The Borel σ-algebra is the standard choice for probability theory on Euclidean spaces. It is rich enough to include practically all sets of interest—intervals, closed sets, countable unions, etc.—while avoiding pathological non-measurable sets.

1.2.4 Probability Measure: Kolmogorov's Axioms

Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a **probability measure** if it satisfies the following **Kolmogorov axioms**:

1. **Non-negativity:** $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.
2. **Normalization:** $\mathbb{P}(\Omega) = 1$.
3. **Countable additivity (σ-additivity):** For any sequence $\{A_n\}_{n=1}^{\infty}$ of pairwise disjoint events in $\mathcal{F}$,

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

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

1.2.5 Elementary Properties of Probability Measures

For any events $A, B \in \mathcal{F}$, the following properties follow directly from the Kolmogorov axioms:

1. **Probability of the empty set:** $\mathbb{P}(\emptyset) = 0$.
2. **Complement rule:** $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
3. **Boundedness:** $0 \leq \mathbb{P}(A) \leq 1$.
4. **Monotonicity:** If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
5. **Finite additivity:** If $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
6. **Inclusion–Exclusion Principle:**

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

1.2.6 Proof: Complement Rule

Theorem: For any $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Proof:

1. By definition of complement, $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$.
2. Since A and A^c are disjoint, finite additivity gives:
$$\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c).$$
3. By normalization, $\mathbb{P}(\Omega) = 1$, hence:
$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1.$$
4. Therefore, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. □

1.2.7 Proof: Inclusion–Exclusion Principle

Theorem: For any $A, B \in \mathcal{F}$,
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Proof:

1. Decompose $A \cup B$ as the disjoint union:
$$A \cup B = A \cup (B \setminus A),$$
with $A \cap (B \setminus A) = \emptyset$.
2. By finite additivity:
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A).$$
3. Similarly, decompose $B = (B \cap A) \cup (B \setminus A)$, a disjoint union:
$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A).$$
4. Thus, $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
5. Substituting back yields:
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$
□

1.3 Conditional Probability

1.3.1 Motivation

Conditional probability addresses the fundamental problem of **updating beliefs in light of partial information**. In scientific practice, we rarely operate in an information vacuum; we typically possess data that restricts the set of possible outcomes. Conditional probability provides the mathematical mechanism to move from a *prior* probability to a *posterior* probability, quantifying how the uncertainty about an event A changes when we learn that event B has occurred.

Geometric Interpretation: When informed that B has occurred, the universe of discourse shrinks: all outcomes outside B become impossible. Geometrically, B becomes the new sample space. The conditional probability $\mathbb{P}(A \mid B)$ is the proportion of the area of A that lies within B relative to the total area of B:
$$\mathbb{P}(A \mid B) = \frac{\text{area}(A \cap B)}{\text{area}(B)}.$$

1.3.2 Definition of Conditional Probability

For fixed $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the **conditional probability** of A given B, denoted $\mathbb{P}(A \mid B)$, is defined as:

$$\mathbb{P}(A \mid B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \forall A \in \mathcal{F}.$$

Intuition: - $\mathbb{P}(B \mid B) = 1$, confirming that B becomes the certain event. - If $A \cap B = \emptyset$, then $\mathbb{P}(A \mid B) = 0$.

Important: For fixed B, the mapping $A \mapsto \mathbb{P}(A \mid B)$ is itself a probability measure on $(\Omega, \mathcal{F})$, satisfying all Kolmogorov axioms.

1.3.3 The Law of Total Probability (LTP)

Let $\{B_n\}_{n=1}^{\infty}$ be a **partition** of Ω satisfying: 1. $B_i \cap B_j = \emptyset$ for $i \neq j$; 2. $\bigcup_{n=1}^{\infty} B_n = \Omega$; 3. $\mathbb{P}(B_n) > 0$ for all n.

Theorem (Law of Total Probability): For any event $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A \mid B_n)\, \mathbb{P}(B_n).$$

Proof:

1. Since $\{B_n\}$ partitions Ω,
$$A = A \cap \Omega = A \cap \left(\bigcup_{n=1}^{\infty} B_n \right) = \bigcup_{n=1}^{\infty} (A \cap B_n).$$
2. The sets $\{A \cap B_n\}_{n=1}^{\infty}$ are pairwise disjoint because the B_n's are.
3. By σ-additivity:
$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A \cap B_n).$$
4. From the definition of conditional probability, $\mathbb{P}(A \cap B_n) = \mathbb{P}(A \mid B_n)\mathbb{P}(B_n)$.
5. Substitution yields the result. □

1.4 Bayes' Theorem

1.4.1 Motivation

Bayes' Theorem solves the problem of **inverse inference**. While classical probability typically addresses forward questions—"Given a cause, what is the probability of observing a particular effect?"—Bayes' Theorem provides the answer to the reverse: **"Given the observed evidence (effect), what is the probability that a specific cause generated it?"** This theorem is the cornerstone of sequential learning, enabling the transformation of a *prior* belief into a *posterior* belief after observing new data.

1.4.2 Bayes' Theorem (Simple Form)

For events $A, B \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

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

Proof:

1. From the definition of conditional probability:

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

2. Also, $\mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$.
3. Substituting gives the stated formula. □

1.4.3 Bayes' Theorem (Partition Form)

Let $\{A_1, A_2, \ldots, A_n\}$ be a partition of Ω with $\mathbb{P}(A_i) > 0$ for all i. For any event B with $\mathbb{P}(B) > 0$:

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

Terminology: - $\mathbb{P}(A_i)$: **Prior probability** (belief before observing data). - $\mathbb{P}(B \mid A_i)$: **Likelihood** (probability of evidence under hypothesis A_i). - $\mathbb{P}(A_i \mid B)$: **Posterior probability** (updated belief after observing B).

1.5 Independence of Events

1.5.1 Motivation

Independence formalizes the **absence of informational influence** between events. In practical terms, two events are independent if knowledge of one provides no information about the likelihood of the other. This concept is essential for simplifying calculations in repeated experiments (e.g., coin tosses, sampling with replacement) and for isolating variables in statistical models.

Geometric Interpretation: Visualize Ω as a unit square. Let A be a vertical strip of width a, so $\mathbb{P}(A) = a$, and B a horizontal strip of height b, so $\mathbb{P}(B) = b$. Their intersection $A \cap B$ is the rectangle formed by the crossing of these strips. Events A and B are **independent** if the area of this rectangle equals the product of the individual areas:

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

Equivalently, the proportion of A within B equals the proportion of A in the entire space Ω.

1.5.2 Independence of Two Events

Two events $A, B \in \mathcal{F}$ are **statistically independent** if and only if:

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

Remarks: - This definition is valid even when $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$. - If $\mathbb{P}(B) > 0$, independence is equivalently characterized by:

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

1.5.3 Mutual Independence of n Events

A collection of events $\{A_1, A_2, \ldots, A_n\}$ is **mutually independent** if, for every non-empty subcollection of indices $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$,

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

Important: Pairwise independence—i.e., $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$—does **not** imply mutual independence.

1.5.4 Properties of Independence

If A and B are independent, then the following pairs are also independent: 1. A and B^c; 2. A^c and B; 3. A^c and B^c.

Proof for A and B^c:

1. Decompose A as the disjoint union:

$$A = (A \cap B) \cup (A \cap B^c).$$

2. By additivity:

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

3. By independence, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Thus:

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)[1 - \mathbb{P}(B)] = \mathbb{P}(A)\mathbb{P}(B^c).$$

□

1.6 Independence of σ-Algebras

1.6.1 Motivation

The extension of independence to σ-algebras represents the definitive step from elementary probability—based on isolated events—to the theory of stochastic processes and asymptotic inference. While independence of events concerns individual occurrences, independence of σ-algebras concerns the **independence of entire information flows**.

Geometric Interpretation in Product Spaces: Consider a sample space $\Omega = \Omega_1 \times \Omega_2$. Let $\mathcal{G}$ describe events in the first coordinate (horizontal axis) and $\mathcal{H}$ describe events in the second (vertical axis). These σ-algebras are independent if the probability of any measurable rectangle $G \times H$ equals the product of the marginal probabilities.

1.6.2 Formal Definition

Let $\{\mathcal{F}_i\}_{i \in I}$ be a collection of sub-σ-algebras of $\mathcal{F}$. They are **independent** if, for any finite subcollection $\{i_1, \ldots, i_k\} \subseteq I$ and any events $A_{i_j} \in \mathcal{F}_{i_j}$,

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

1.6.3 Independence of Random Variables

Random variables $X_1, \ldots, X_m$ are **independent** if their generated σ-algebras $\sigma(X_1), \ldots, \sigma(X_m)$ are independent.

Equivalent Characterization: For any Borel sets $B_1, \ldots, B_m \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X_1 \in B_1, \ldots, X_m \in B_m) = \prod_{j=1}^{m} \mathbb{P}(X_j \in B_j).$$

Consequences of Independence: - If X and Y are independent, then for any measurable functions $g, h : \mathbb{R} \to \mathbb{R}$, the variables $g(X)$ and $h(Y)$ are also independent. - In particular (provided expectations exist),

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

1.7 Summary of Key Theorems

Theorem	Formula
Complement Rule	$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
Inclusion–Exclusion	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
Law of Total Probability	$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$
Bayes' Theorem	$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B\|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B\|A_j)\mathbb{P}(A_j)}$
Independence (Events)	$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$
Independence (σ-Algebras)	$\mathbb{P}(\cap_j A_{i_j}) = \prod_j \mathbb{P}(A_{i_j})$

Chapter 2: Random Variables and Distributions

2.1 Definition of Random Variable

2.1.1 Motivation

The fundamental problem addressed in this chapter is the **numerical quantification of qualitative phenomena**. Sample spaces may contain abstract outcomes—such as "Success/Failure," "Heads/Tails," or states of an economic system—on which we cannot directly apply tools from calculus, linear algebra, or real analysis. The random variable serves as a translator that maps uncertainty from the abstract space Ω to the real line $\mathbb{R}$, enabling the use of measures of central tendency, dispersion, and other statistical summaries.

Geometric Interpretation: Imagine the sample space Ω as a cloud of abstract points. The random variable X acts as a projection that "pushes" each outcome $\omega \in \Omega$ to a specific location on the real line. An event of interest—such as "profit is positive"—corresponds to an interval or Borel set in $\mathbb{R}$, and its probability is computed as the measure of the preimage of that set under X.

2.1.2 Formal Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a **(real-valued) random variable** if it is **measurable** with respect to the Borel σ-algebra on $\mathbb{R}$: for every Borel set $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) \equiv \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}.$$

Equivalently, X is a random variable if and only if

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}, \quad \forall x \in \mathbb{R}.$$

This equivalence follows from the fact that the collection of intervals $(-\infty, x]$ generates the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

2.1.3 Measurability Criterion

Theorem: A function $X : \Omega \to \mathbb{R}$ is a random variable if and only if $X^{-1}((-\infty, x]) \in \mathcal{F}$ for all $x \in \mathbb{R}$.

Proof:

1. Define the collection
$$\mathcal{M} = \{B \subseteq \mathbb{R} \mid X^{-1}(B) \in \mathcal{F}\}.$$

2. We first show that $\mathcal{M}$ is a σ-algebra:
   - $\mathbb{R} \in \mathcal{M}$, since $X^{-1}(\mathbb{R}) = \Omega \in \mathcal{F}$.
   - If $B \in \mathcal{M}$, then $X^{-1}(B^c) = [X^{-1}(B)]^c \in \mathcal{F}$, so $B^c \in \mathcal{M}$.
   - If $\{B_n\}_{n=1}^{\infty} \subseteq \mathcal{M}$, then
   $$X^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right) = \bigcup_{n=1}^{\infty} X^{-1}(B_n) \in \mathcal{F},$$
   so $\bigcup_{n=1}^{\infty} B_n \in \mathcal{M}$.

3. Since $\mathcal{M}$ contains all intervals of the form $(-\infty, x]$, and these intervals generate $\mathcal{B}(\mathbb{R})$, we have $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{M}$.

4. Therefore, $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$, establishing measurability. □

2.2 Induced Distribution (Push-forward Measure)

2.2.1 Motivation

The induced distribution addresses the problem of **rendering the model independent of the original sample space**. By transporting the entire probabilistic structure from Ω to $\mathbb{R}$, we obtain a new probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ where standard analytical tools become applicable. This construction is the foundation for working exclusively with distributions, abstracting away the underlying probability space.

Geometric Interpretation: Imagine the probability measure $\mathbb{P}$ as a unit mass distributed over Ω. The random variable X acts as a conveyor belt that moves this mass from Ω to the real line. The induced distribution $\mathbb{P}_X$ is the "shadow" or "projection" of the original probability onto the numerical codomain.

2.2.2 Formal Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. The **induced probability measure** (or **push-forward measure**) of X, denoted $\mathbb{P}_X$ or $\mathbb{P} \circ X^{-1}$, is the function $\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$ defined by
$$\mathbb{P}_X(B) \equiv \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}), \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

2.2.3 Properties of the Induced Distribution

Theorem: $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof:

1. **Non-negativity:** For any $B \in \mathcal{B}(\mathbb{R})$,
$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) \geq 0,$$
since $\mathbb{P}$ is a probability measure.

2. **Normalization:**
$$\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(X^{-1}(\mathbb{R})) = \mathbb{P}(\Omega) = 1.$$

3. **Countable additivity:** Let $\{B_n\}_{n=1}^{\infty}$ be a sequence of pairwise disjoint Borel sets. Then $\{X^{-1}(B_n)\}_{n=1}^{\infty}$ are pairwise disjoint subsets of Ω, and
$$\mathbb{P}_X\left(\bigcup_{n=1}^{\infty} B_n\right) = \mathbb{P}\left(X^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right)\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} X^{-1}(B_n)\right) = \sum_{n=1}^{\infty} \mathbb{P}(X^{-1}(B_n)) = \sum_{n=1}^{\infty} \mathbb{P}_X(B_n).$$

□

2.3 Cumulative Distribution Function (CDF)

2.3.1 Motivation

The CDF provides a **standardized, complete description** of a random variable's distribution. Prior to the CDF, discrete variables required sums (probability mass functions) and continuous variables required integrals (probability density functions). The CDF unifies both cases as a single function that accumulates probability mass from $-\infty$ to any point $x \in \mathbb{R}$.

Geometric Interpretation: If we imagine the probability measure as a unit mass distributed over the real line, the CDF $F_X(x)$ represents the total mass accumulated from $-\infty$ up to the point x.

2.3.2 Formal Definition

Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The **cumulative distribution function** (CDF) of X, denoted $F_X : \mathbb{R} \to [0, 1]$, is defined by

$$F_X(x) \equiv \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \leq x\}), \quad \forall x \in \mathbb{R}.$$

Equivalently, $F_X(x) = \mathbb{P}_X((-\infty, x])$.

2.3.3 Fundamental Properties of the CDF

Every CDF satisfies the following properties:

1. **Monotonicity:** If $x_1 \leq x_2$, then $F_X(x_1) \leq F_X(x_2)$.
2. **Limits at infinity:**
$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1.$$
3. **Right-continuity:** For every $x \in \mathbb{R}$,
$$\lim_{h \downarrow 0} F_X(x + h) = F_X(x).$$
4. **Interval probability:** For any $a < b$,
$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

2.3.4 Proof of the Interval Probability Formula

Proof:

1. Decompose the event $\{X \leq b\}$ as the disjoint union
$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}.$$
2. By finite additivity of $\mathbb{P}$,
$$\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b).$$
3. By the definition of the CDF,
$$F_X(b) = F_X(a) + \mathbb{P}(a < X \leq b).$$

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

□

2.4 Probability Density Function (PDF)

2.4.1 Motivation

The PDF addresses the problem of **assigning probability to uncountable sets**. For a continuous random variable X, the probability of observing any exact value is zero—$\mathbb{P}(X = x) = 0$. Without the PDF, we would be unable to distinguish regions of higher or lower probability concentration along the real line.

Geometric Interpretation: Visualize the PDF $f_X(x)$ as a non-negative curve above the horizontal axis. The probability of an event is not given by the height of the curve, but by the **area under the curve** over the corresponding interval. The PDF thus functions analogously to a mass density in classical physics.

2.4.2 Formal Definition

Let X be a continuous random variable with CDF F_X. A function $f_X : \mathbb{R} \to \mathbb{R}_+$ is the **probability density function** (PDF) of X if it satisfies:

1. **Non-negativity:** $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.
2. **Normalization:**
$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$
3. **Relationship with the CDF:** For all $x \in \mathbb{R}$,
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

2.4.3 Relationship between PDF and CDF

If F_X is absolutely continuous (and hence differentiable almost everywhere), then

$$f_X(x) = \frac{d}{dx} F_X(x)$$

at every point where F_X is differentiable.

Proof:

1. By definition, $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$.
2. By the Fundamental Theorem of Calculus (applied to absolutely continuous functions),
$$\frac{d}{dx} F_X(x) = f_X(x)$$
for almost every $x \in \mathbb{R}$. □

2.4.4 Probability Calculation via PDF

For any Borel set $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,dx.$$

In particular, for any interval $[a, b]$,

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

2.5 Probability Mass Function (PMF)

2.5.1 Motivation

The PMF addresses the problem of **assigning specific weights to isolated outcomes**. While in continuous spaces the probability of any exact point is zero, discrete phenomena—such as the number of successes in a sequence of Bernoulli trials—exhibit probability concentrated on countable "atoms."

Geometric Interpretation: The PMF is visualized as a **bar chart**. Unlike the PDF, where probability corresponds to area, in the PMF the probability of a point x_i is the **height** of the bar at that point.

2.5.2 Formal Definition

Let X be a discrete random variable with countable support $\mathcal{X} = \{x_1, x_2, \dots\} \subset \mathbb{R}$. The **probability mass function** (PMF) of X, denoted $p_X(x)$, is defined by

$$p_X(x) \equiv \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}).$$

In measure-theoretic terms, the PMF is the Radon–Nikodým derivative of $\mathbb{P}_X$ with respect to the counting measure μ_c:

$$p_X(x) = \frac{d\mathbb{P}_X}{d\mu_c}(x).$$

2.5.3 Properties of the PMF

1. **Non-negativity:** $p_X(x) \ge 0$ for all $x \in \mathcal{X}$.
2. **Normalization:**
$$\sum_{x \in \mathcal{X}} p_X(x) = 1.$$
3. **Relationship with the CDF:** For any $x \in \mathbb{R}$,
$$F_X(x) = \sum_{x_i \le x} p_X(x_i).$$
4. **Probability of any event:** For any subset $B \subseteq \mathcal{X}$,
$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x).$$

2.5.4 Proof of Normalization

Proof:

1. Define the events $A_i = \{\omega \in \Omega \mid X(\omega) = x_i\}$ for $i \in \mathbb{N}$.
2. The collection $\{A_i\}_{i=1}^{\infty}$ forms a partition of $\{\omega \mid X(\omega) \in \mathcal{X}\}$.
3. Since X is discrete, $\mathbb{P}(X \in \mathcal{X}) = 1$.
4. By countable additivity,

$$1 = \mathbb{P}(X \in \mathcal{X}) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) = \sum_{x \in \mathcal{X}} p_X(x).$$

□

2.6 Transformations of Random Variables

2.6.1 Motivation

The transformation problem is central to statistical modeling: we often observe a quantity X, but the theoretical model operates on a transformed quantity $Y = g(X)$. We require a rigorous method to derive the distribution of Y from the known distribution of X.

Geometric Interpretation: A monotone transformation g "stretches" or "compresses" portions of the real line. If g stretches an interval, the probability mass in that region must spread out, resulting in a lower density—the **Jacobian** factor compensates for this local change in "volume."

2.6.2 Transformation Formula (Univariate, Monotone Case)

Let X be a continuous random variable with PDF f_X and support $\mathcal{X} \subseteq \mathbb{R}$. Let $Y = g(X)$, where $g : \mathcal{X} \to \mathbb{R}$ is strictly monotone and continuously differentiable. Then the PDF of Y is given by

$$\boxed{f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|}$$

for $y \in g(\mathcal{X})$, and $f_Y(y) = 0$ otherwise.

2.6.3 Proof for the Strictly Increasing Case

Proof:

1. By definition,

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

2. Since g is strictly increasing,

$$g(X) \leq y \iff X \leq g^{-1}(y).$$

3. Therefore,

$$F_Y(y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

4. Differentiating with respect to y (using the chain rule),

$$f_Y(y) = F_Y'(y) = F_X'(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) = f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y).$$

5. Since g^{-1} is increasing, its derivative is non-negative, so

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|.$$

□

2.6.4 Important Special Cases

1. Linear Transformation: If $Y = aX + b$ with $a \neq 0$, then

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

2. Quadratic Transformation ($Y = X^2$): For $y > 0$,

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left[f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right].$$

3. Multivariate Transformation: Let $g : \mathbb{R}^m \to \mathbb{R}^m$ be a diffeomorphism. Then

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |\det(J_{g^{-1}}(y))|,$$

where $J_{g^{-1}}$ denotes the Jacobian matrix of the inverse transformation.

2.7 Probability Integral Transformation (PIT)

2.7.1 Motivation

The Probability Integral Transformation (PIT) solves two interrelated problems: **standardization of uncertainties** and **simulation of arbitrary distributions**. The PIT establishes that any continuous random variable can be transformed into a uniform variable $U \sim \text{Unif}(0, 1)$, and conversely, any uniform variable can be transformed into a variable with any desired CDF.

Geometric Interpretation: The PIT "flattens" any complex probability density into a rectangle of unit height over the interval $[0, 1]$. This is analogous to "unfolding" the distribution.

2.7.2 Formal Statement

Let X be a continuous random variable with CDF F_X. Define

$$U = F_X(X).$$

Theorem (Probability Integral Transform): If F_X is continuous, then

$$U \sim \text{Unif}(0, 1).$$

2.7.3 Proof

Proof:

1. For $u \in [0, 1]$,

$$F_U(u) = \mathbb{P}(U \leq u) = \mathbb{P}(F_X(X) \leq u).$$

2. Since F_X is continuous and non-decreasing,

$$F_X(X) \leq u \iff X \leq F_X^{-1}(u),$$

where $F_X^{-1}(u) = \inf\{x \mid F_X(x) \geq u\}$ is the generalized inverse.

3. Therefore,

$$F_U(u) = \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u.$$

4. Thus, $F_U(u) = u$ for $u \in [0,1]$, which is precisely the CDF of the $\text{Unif}(0,1)$ distribution.

5. The corresponding density is

$$f_U(u) = \frac{d}{du} F_U(u) = 1, \quad u \in [0,1].$$

□

2.7.4 Converse: Generating Samples

Given $U \sim \text{Unif}(0,1)$, the variable

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

has CDF F_X. This is the standard method for generating random samples from any continuous distribution.

2.8 The σ-Algebra Generated by a Random Variable

2.8.1 Motivation

The σ-algebra generated by X formalizes the **quantification of available information**. In complex systems, we do not observe the full state $\omega \in \Omega$, but only functions such as asset prices or sensor readings. The σ-algebra $\sigma(X)$ defines precisely which questions about the experiment can be answered given knowledge of X.

Geometric Interpretation: If two outcomes ω_1 and ω_2 yield the same value $X(\omega_1) = X(\omega_2)$, they are indistinguishable under $\sigma(X)$. Thus, $\sigma(X)$ partitions Ω into equivalence classes where X is constant.

2.8.2 Formal Definition

Let $X : \Omega \to \mathbb{R}$ be a random variable. The **σ-algebra generated by** X, denoted $\sigma(X)$, is the collection of subsets of Ω defined by

$$\sigma(X) \equiv \{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}.$$

By construction, $\sigma(X)$ is the **smallest** σ-algebra on Ω that makes X a measurable function.

2.8.3 Fundamental Properties

1. **Subordination:** $\sigma(X) \subseteq \mathcal{F}$, since X is measurable with respect to $\mathcal{F}$.
2. **Atomic structure:** $\sigma(X)$ contains all sets of the form $\{X \leq x\}$, $\{X = x\}$, $\{X \in [a,b]\}$, etc.
3. **Monotonicity under composition:** If $Y = g(X)$ for a measurable function $g : \mathbb{R} \to \mathbb{R}$, then

$$\sigma(Y) \subseteq \sigma(X).$$

4. **Equality condition:** If g is bijective with measurable inverse, then $\sigma(Y) = \sigma(X)$.

2.9 The Radon–Nikodým Theorem

2.9.1 Motivation

The Radon–Nikodým theorem provides the theoretical foundation for **representing measures as integrals of functions**. It guarantees that, under appropriate compatibility conditions, any measure can be expressed as the integral of a density function with respect to a reference measure. This theorem justifies the existence of PDFs and PMFs as special cases.

Geometric Interpretation: Imagine two mass distributions on a metal plate—ν and μ. If the mass ν is "absolutely continuous" with respect to μ (i.e., wherever μ has no mass, ν has no mass), then there exists a "relative density" function f. Multiplying the local mass of μ by $f(x)$ reconstructs exactly the mass of ν.

2.9.2 Formal Statement

Let $(X, \mathcal{A})$ be a measurable space. Let ν be a σ-finite signed measure and μ a σ-finite measure on $(X, \mathcal{A})$.

We say that ν is **absolutely continuous** with respect to μ, denoted $\nu \ll \mu$, if

$$\forall A \in \mathcal{A}, \quad \mu(A) = 0 \implies \nu(A) = 0.$$

Theorem (Radon–Nikodým): If $\nu \ll \mu$, then there exists a measurable function $f : X \to [0, \infty)$ such that

$$\nu(A) = \int_A f \, d\mu, \quad \forall A \in \mathcal{A}.$$

The function f is unique μ-almost surely and is called the **Radon–Nikodým derivative**, denoted

$$f = \frac{d\nu}{d\mu}.$$

2.9.3 Applications in Probability Theory

1. **Continuous random variables:** The PDF is the Radon–Nikodým derivative of $\mathbb{P}_X$ with respect to the Lebesgue measure λ:
$$f_X(x) = \frac{d\mathbb{P}_X}{d\lambda}(x).$$

2. **Discrete random variables:** The PMF is the Radon–Nikodým derivative of $\mathbb{P}_X$ with respect to the counting measure μ_c:
$$p_X(x) = \frac{d\mathbb{P}_X}{d\mu_c}(x).$$

3. **Conditional expectation:** For a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, the conditional expectation $\mathbb{E}[Y \mid \mathcal{G}]$ is the Radon–Nikodým derivative of the measure
$$\nu(B) = \int_B Y \, d\mathbb{P}, \quad B \in \mathcal{G},$$
with respect to $\mathbb{P}$ restricted to $\mathcal{G}$.

2.10 Summary of Key Results

Concept	Definition	Key Formula
Random Variable	Measurable function $X : \Omega \to \mathbb{R}$	$X^{-1}(B) \in \mathcal{F},\ \forall B \in \mathcal{B}(\mathbb{R})$
Induced Distribution	$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B))$	$\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$
CDF	$F_X(x) = \mathbb{P}(X \le x)$	$\mathbb{P}(a < X \le b) = F_X(b) - F_X(a)$
PDF	$f_X(x) = \frac{d}{dx} F_X(x)$	$\mathbb{P}(a \le X \le b) = \int_a^b f_X(x)\,dx$
PMF	$p_X(x) = \mathbb{P}(X = x)$	$\sum_{x \in \mathcal{X}} p_X(x) = 1$
Transformation	$Y = g(X)$	$f_Y(y) = f_X(g^{-1}(y)) \cdot \left\lvert \frac{d}{dy} g^{-1}(y) \right\rvert$
PIT	$U = F_X(X)$	$U \sim \text{Unif}(0,1)$
$\sigma(X)$	$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$	Smallest σ-algebra making X measurable
Radon–Nikodým	$f = \frac{d\nu}{d\mu}$	$\nu(A) = \int_A f\,d\mu$ iff $\nu \ll \mu$

Chapter 3: Probability Distributions

3.1 Bernoulli Distribution

3.1.1 Motivation and Context

The Bernoulli distribution addresses the fundamental problem of **modeling dichotomous experiments**—situations with exactly two possible outcomes. By convention, these outcomes are labeled as **"success"** (coded as 1) and **"failure"** (coded as 0).

In economics and social sciences, this model is ubiquitous for describing binary choice phenomena or binary states of nature. **Typical examples** include: whether an airline passenger shows up for boarding, whether an individual possesses a college degree, whether a coin toss results in "heads," or whether a manufactured component is defective.

Geometric Interpretation: The distribution can be visualized as **two impulses** (or vertical bars) located at points 0 and 1 on the real line, with heights corresponding to the respective probabilities $1 - p$ and p.

3.1.2 Formal Definition

Let X be a discrete random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that X follows a **Bernoulli distribution** with parameter $p \in [0,1]$, denoted $X \sim \text{Bernoulli}(p)$, if its probability mass function (PMF) is

$$\boxed{p_X(x) = p^x (1-p)^{1-x}, \quad x \in \{0,1\}}.$$

Equivalently,

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

Remark: The parameter p is assumed constant throughout the experiment. If p varied across trials, the Bernoulli structure would no longer hold.

3.1.3 Statistical Properties

Property	Formula
Mean	$\mathbb{E}[X] = p$
Variance	$\mathrm{Var}(X) = p(1-p)$
MGF	$M_X(t) = 1 - p + pe^t$
Support	$\{0, 1\}$

3.1.4 Proof of the Moments

Mean:

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

Second moment:

$$\mathbb{E}[X^2] = 0^2 \cdot \mathbb{P}(X = 0) + 1^2 \cdot \mathbb{P}(X = 1) = p.$$

Variance:

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

□

3.1.5 Key Connections

- **Sum of Bernoullis:** The sum of n independent and identically distributed (i.i.d.) Bernoulli(p) random variables follows a **Binomial distribution** with parameters n and p.
- **Foundation for Binary Choice Models:** The Bernoulli distribution is the building block for **Logit** and **Probit** models in econometrics, where the dependent variable is binary and the likelihood is constructed assuming each observation follows a Bernoulli distribution conditional on covariates.

3.2 Binomial Distribution

3.2.1 Motivation and Context

The Binomial distribution generalizes the Bernoulli distribution from a single trial to a fixed number of independent trials. It models the **total number of successes** in n identical, independent Bernoulli experiments.

Binomial Experiment Conditions:

1. Consists of n **Bernoulli trials** (each with two possible outcomes).
2. The trials are **statistically independent**.
3. The success probability p is **constant** across all trials.

A Binomial random variable X has support $\mathcal{X} = \{0, 1, 2, \ldots, n\}$.

Geometric Behavior: The PMF is unimodal, increasing monotonically to a maximum and then decreasing. The mode occurs at the largest integer $k \leq (n+1)p$.

3.2.2 Probability Mass Function

Let $X \sim \text{Binomial}(n, p)$. The probability of observing exactly k successes is

$$\boxed{\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n},$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

3.2.3 Statistical Properties

Property	Formula
Mean	$\mathbb{E}[X] = np$
Variance	$\text{Var}(X) = np(1-p)$
MGF	$M_X(t) = (1-p+pe^t)^n$
Mode	Largest integer $\leq (n+1)p$

3.2.4 Key Approximations

1. **Poisson Approximation:** When $n \to \infty$ and $p \to 0$ such that $np = \lambda$ remains constant, the Binomial converges to a **Poisson distribution** with parameter λ. This is useful for modeling "rare events" in large populations. **Rule of thumb:** $n \geq 20$ and $p \leq 0.05$.
2. **Normal Approximation (De Moivre–Laplace):** For large n, the standardized variable

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1).$$

   Rule of thumb: $np > 5$ and $n(1-p) > 5$.
3. **Hypergeometric Connection:** The Hypergeometric distribution (sampling without replacement) can be approximated by the Binomial when the population size is much larger than the sample size (typically $n/N \leq 0.10$).

3.2.5 Applications in Econometrics

- **Binary Choice Models:** The likelihood function for models such as **Logit** and **Probit** is constructed from Bernoulli contributions (Binomial with $n = 1$).
- **Count Data with Upper Bound:** Binomial regression is used when the dependent variable is a count of successes with a known upper limit n_i (e.g., number of children graduating in a family of n_i children).
- **Proportion Estimation:** The MLE for the population proportion p is the sample proportion $\hat{p} = X/n$, which is unbiased and has minimum variance.

3.3 Poisson Distribution

3.3.1 Motivation and Context

The Poisson distribution models the **number of occurrences of events** in a fixed interval of time, space, volume, or surface. It is characterized by a single parameter $\lambda > 0$, which equals both the mean and the variance.

Poisson Process Assumptions:

1. **Independence:** The numbers of events in disjoint intervals are independent.
2. **Constant rate:** The probability of an occurrence is proportional to the interval length.
3. **Singularity:** The probability of more than one event in an infinitesimally small interval is negligible.

Typical Examples: Number of calls received by a call center, equipment failures, traffic accidents, orders received by a company, or radioactive particle decays.

3.3.2 Probability Mass Function

Let $X \sim \text{Poisson}(\lambda)$ with $\lambda > 0$. The PMF is

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

3.3.3 Statistical Properties

Property	Formula
Mean	$\mathbb{E}[X] = \lambda$
Variance	$\text{Var}(X) = \lambda$
MGF	$M_X(t) = \exp\{\lambda(e^t - 1)\}$
Support	$\{0, 1, 2, \dots\}$

Key Feature (Equidispersion): The mean equals the variance—a characteristic that can be restrictive in practice.

Reproductive Property: The sum of independent Poisson variables is Poisson with parameter equal to the sum of the individual parameters:

$$X_i \sim \text{Poisson}(\lambda_i) \text{ independent } \implies \sum_{i=1}^{n} X_i \sim \text{Poisson}\left(\sum_{i=1}^{n} \lambda_i\right).$$

3.3.4 Relationships with Other Distributions

1. **Limit of Binomial:** $\text{Binomial}(n, p) \to \text{Poisson}(\lambda)$ when $n \to \infty$, $p \to 0$, and $np = \lambda$ remains constant. **Rule of thumb:** $n \geq 20$ and $p \leq 0.05$.
2. **Relationship with Exponential:** In a Poisson process with rate λ, the **waiting time** between two consecutive events follows an **Exponential distribution** with parameter λ.
3. **Relationship with Gamma:** The total waiting time until the r-th event follows a **Gamma distribution** with shape r and rate λ.

3.3.5 Poisson Regression in Econometrics

Poisson regression is the standard model for **count data** in econometrics.

- **Conditional mean:** Modeled as
$$\mathbb{E}[Y \mid X] = \exp(X'\beta),$$
ensuring positivity of the predicted mean.
- **Estimation (MLE):** The log-likelihood is globally concave, making optimization computationally efficient.
- **QMLE Robustness:** The Poisson estimator is a **Quasi-Maximum Likelihood Estimator (QMLE)**—it remains consistent for the mean parameters even if the true distribution is not Poisson, as long as the conditional mean is correctly specified.
- **Overdispersion:** When $\text{Var}(Y \mid X) > \mathbb{E}[Y \mid X]$, the data exhibit overdispersion. In such cases, the **Negative Binomial** model is preferred, or robust (sandwich) standard errors should be used.

3.4 Normal Distribution

3.4.1 Definition and Functional Form

The **Normal distribution** (or **Gaussian distribution**) is the most important probability model in statistics and econometrics, serving as the cornerstone for inference, error modeling, and asymptotic theory.

A continuous random variable X follows a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, denoted $X \sim N(\mu, \sigma^2)$, if its PDF is

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

Standard Normal: When $\mu = 0$ and $\sigma^2 = 1$, we have the **standard normal** distribution, denoted $Z \sim N(0, 1)$, with PDF

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad z \in \mathbb{R},$$

and CDF $\Phi(z) = \int_{-\infty}^{z} \phi(t)\, dt$.

3.4.2 Properties and Geometry

Property	Description
Shape	Bell-shaped, symmetric about μ
Mean = Median = Mode	All equal and located at μ
Empirical Rule	68–95–99.7 rule: $\mu \pm \sigma$ (68%), $\mu \pm 2\sigma$ (95%), $\mu \pm 3\sigma$ (99.7%)
Skewness	0 (symmetric)
Excess Kurtosis	0 (mesokurtic, kurtosis = 3)
Inflection Points	At $x = \mu \pm \sigma$

Key Property: For jointly normal variables, **zero covariance implies independence**—a property not shared by other distribution families.

3.4.3 Operations and Transformations

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

2. **Linear Combinations:** Any affine transformation of a normal variable is normal:
$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

3. **Sum of Independent Normals:** If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then
$$\sum_{i=1}^{n} X_i \sim N\left(\sum \mu_i, \sum \sigma_i^2\right).$$

4. **Derived Distributions:**
   - $Z^2 \sim \chi_1^2$
   - Ratios of normals give rise to the t and F distributions.

3.4.4 Role in Econometrics

Central Limit Theorem (CLT): The sum (or mean) of a large number of i.i.d. variables with finite variance **converges in distribution to a normal**, regardless of the original distribution. This provides the theoretical basis for assuming normality of regression errors asymptotically.

Classical Normal Linear Regression Model (CNLRM): Under $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$:

1. OLS estimators are **Best Unbiased Estimators (BUE)** and have **exact normal distributions** in finite samples.
2. Hypothesis tests (t and F) and confidence intervals are valid without relying on asymptotic arguments.

3.4.5 Testing for Normality

The **Jarque–Bera (JB) test** is the standard test for normality of residuals, jointly testing whether skewness is 0 and excess kurtosis is 0:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \xrightarrow{d} \chi^2_2,$$

where S is the sample skewness and K is the sample kurtosis.

If normality fails in small samples, t and F tests lose finite-sample validity, although in large samples they remain valid due to **asymptotic normality**.

3.5 Exponential Distribution

3.5.1 Definition and Properties

The **Exponential distribution** is a continuous distribution with support on $[0, \infty)$, widely used for modeling waiting times, lifetimes, and durations.

PDF (Rate parameterization):

$$f_X(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

where $\lambda > 0$ is the **rate parameter**.

PDF (Scale parameterization):

$$f_X(x; \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x \geq 0,$$

where $\beta = 1/\lambda > 0$ is the **scale parameter** (mean).

Property	Formula
Mean	$\mathbb{E}[X] = 1/\lambda = \beta$
Variance	$\mathrm{Var}(X) = 1/\lambda^2 = \beta^2$
CDF	$F_X(x) = 1 - e^{-\lambda x}$
MGF	$M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$

3.5.2 Memoryless Property

The exponential distribution is characterized by the **memoryless property**: for any $s, t > 0$,

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

This implies that the probability of future survival is independent of time already elapsed—the object is "as good as new" as long as it is functioning. **This property is unique to the Exponential distribution among continuous distributions.**

3.5.3 Relationships with Other Distributions

Relationship	Description
Poisson Process	Waiting time between events follows Exponential(λ)
Gamma	Sum of n i.i.d. Exponential(λ) variables is Gamma(n, λ)
Chi-Square	χ^2_2 is Exponential with $\beta = 2$ (i.e., $\lambda = 1/2$)
Geometric	Exponential is the continuous analog of the Geometric distribution

---

3.5.4 Applications in Econometrics

- **Exponential Trend:** Time series such as GDP or population often exhibit constant proportional growth, modeled as $y_t = \exp(\beta_0 + \beta_1 t + \varepsilon_t)$.
- **Exponential Regression:** Conditional mean $\mathbb{E}[Y \mid X] = \exp(X'\beta)$ ensures positive predictions—standard for **Poisson Regression**.
- **Exponential Smoothing (EWMA):** A forecasting technique where weights decrease exponentially with the age of data.
- **EGARCH:** A volatility model using the exponential form to capture asymmetric effects (the "leverage effect") without imposing non-negativity restrictions.

3.6 Chi-Square Distribution (χ^2)

3.6.1 Definition and Construction

Let $Z_1, Z_2, \ldots, Z_k$ be i.i.d. standard normal variables. The sum of their squares

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k,$$

where $k \in \mathbb{N}$ is the **degrees of freedom**.

PDF:

$$f_X(x; k) = \frac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad x \geq 0.$$

Geometric Shape: The Chi-Square distribution is defined only for $x \geq 0$ and is **right-skewed** (positive skewness), though it becomes more symmetric as k increases.

3.6.2 Statistical Properties

Property	Formula
Mean	$\mathbb{E}[X] = k$
Variance	$\text{Var}(X) = 2k$
MGF	$M_X(t) = (1 - 2t)^{-k/2}$ for $t < 1/2$
Gamma Connection	$\chi^2_k = \text{Gamma}(k/2, 1/2)$

Reproductive Property: If $Q_1 \sim \chi^2_{k_1}$ and $Q_2 \sim \chi^2_{k_2}$ are independent, then

$$Q_1 + Q_2 \sim \chi^2_{k_1+k_2}.$$

3.6.3 Relationships with Other Distributions

1. **Normal:** For large k, $\sqrt{2\chi^2_k} \approx N(\sqrt{2k-1}, 1)$ (Wilson–Hilferty approximation).
2. t **Distribution:** If $Z \sim N(0,1)$ and $V \sim \chi^2_k$ are independent, then

$$\frac{Z}{\sqrt{V/k}} \sim t_k.$$

3. F **Distribution:** If $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ are independent, then

$$\frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2}.$$

3.6.4 Applications in Econometrics

- **Variance Inference:** For normal data, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, enabling **confidence intervals for variance**.
- **Goodness-of-Fit Tests:** Pearson's statistic $\sum(O_i - E_i)^2/E_i$ follows a χ^2 distribution for independence and fit tests.
- **Wald Test:** Quadratic forms of estimators converge to χ^2 with degrees of freedom equal to the number of restrictions.
- **LM Tests:** Tests for heteroskedasticity (Breusch–Pagan/White) and autocorrelation (Breusch–Godfrey) use $nR^2 \sim \chi^2$.
- **Sargan–Hansen Test (GMM):** Tests for overidentifying restrictions in IV/GMM models.

3.7 Student's t Distribution

3.7.1 Definition and Construction

The Student's t distribution with ν degrees of freedom is defined as the distribution of

$$T = \frac{Z}{\sqrt{V/\nu}},$$

where $Z \sim N(0,1)$, $V \sim \chi^2_\nu$, and Z and V are independent.

PDF:

$$f_T(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad t \in \mathbb{R}.$$

3.7.2 Properties

Property	Description
Shape	Symmetric bell-shaped about 0
Tails	Heavier than Normal ("fat tails")—more probability for extreme values
Mean	0 for $\nu > 1$ (undefined otherwise)
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$ (infinite otherwise)
Asymptotic	$t_\nu \to N(0,1)$ as $\nu \to \infty$
MGF	Does not exist (tails too heavy)

3.7.3 Applications

The t distribution is central to inference in the **Classical Normal Linear Regression Model**:

$$T = \frac{\hat{\beta}_j - \beta_{H_0}}{\text{se}(\hat{\beta}_j)} \sim t_{n-k}.$$

Relationship with F: If $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$.

Robustness: The t test is robust to moderate deviations from normality, especially in large samples.

Historical Note: Discovered by **William Sealy Gosset** in 1908 and published under the pseudonym **"Student"** while working at the Guinness brewery.

3.8 Snedecor's F Distribution

3.8.1 Definition and Construction

Let $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ be independent. Then

$$F = \frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2},$$

where ν_1 (numerator) and ν_2 (denominator) are the degrees of freedom.

PDF:

$$f_F(x;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+\frac{\nu_1}{\nu_2}x\right)^{(\nu_1+\nu_2)/2}}, \quad x \geq 0.$$

Support: $x \geq 0$, with right-skewed density.

3.8.2 Statistical Properties

Property	Formula
Mean	$\frac{\nu_2}{\nu_2-2}$ for $\nu_2 > 2$
Variance	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$

3.8.3 Key Relationships

1. t **and** F**:** If $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$.
2. **Inverse Property:** If $F \sim F_{\nu_1,\nu_2}$, then $1/F \sim F_{\nu_2,\nu_1}$.
3. **Limit to χ^2:** As $\nu_2 \to \infty$, $\nu_1 F \xrightarrow{d} \chi^2_{\nu_1}$.

4. **Limit to Normal:** As both $\nu_1, \nu_2 \to \infty$, F approaches a normal distribution.

3.8.4 Applications in Econometrics

- **Linear Restrictions:** Testing $H_0 : R\beta = r$ by comparing the SSR of restricted vs. unrestricted models.
- **Global Significance:** Testing that all slope coefficients are zero:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}.$$

- **Chow Test:** Detecting structural breaks in time series.
- **Equality of Variances:** The ratio of sample variances from two independent normal populations follows an F distribution.
- **ANOVA:** Analysis of variance, comparing means of multiple groups simultaneously.

3.9 Summary Table: Continuous Distributions

Distribution	PDF	Mean	Variance
Uniform(a, b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
χ^2_k	$\frac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)}$	k	$2k$
t_ν	$\frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$	0 (for $\nu > 1$)	$\frac{\nu}{\nu-2}$ (for $\nu > 2$)
F_{ν_1,ν_2}	$\frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+\frac{\nu_1}{\nu_2}x\right)^{(\nu_1+\nu_2)/2}}$	$\frac{\nu_2}{\nu_2-2}$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

Chapter 4: Expectation, Variance, and Moments

4.1 Mathematical Expectation via the Lebesgue Integral

4.1.1 Motivation

The expectation operator addresses the fundamental problem of **quantifying central tendency** in a probability space. While elementary treatments handle discrete cases via sums and continuous cases via Riemann integrals separately, the Lebesgue integral provides a unified framework that applies to *any* random variable defined on a measure space, including mixtures and variables with pathological behavior.

Geometric Interpretation: Imagine the probability measure $\mathbb{P}$ as a unit mass distributed over Ω. The expectation $\mathbb{E}[X]$ is the **center of mass** or **balance point** of this distribution. Just as the center of mass of a physical object is the point where it would perfectly balance, the expectation summarizes the "location" of the probability distribution.

4.1.2 Formal Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The **expectation** (or **expected value**) of a random variable $X : \Omega \to \mathbb{R}$ is defined as the Lebesgue integral of X with respect to the probability measure $\mathbb{P}$:

$$\boxed{\mathbb{E}[X] \equiv \int_\Omega X(\omega)\, d\mathbb{P}(\omega)}.$$

By the **change-of-variables theorem** (a consequence of the Radon–Nikodým theorem), this integral can be transported to the real line using the induced distribution $\mathbb{P}_X$:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}_X(x).$$

If $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure, with density f_X, this reduces to the familiar formula:

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

If X is discrete with PMF p_X, we have:

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

4.1.3 Construction of the Lebesgue Integral

The Lebesgue integral is constructed in three rigorous steps:

Step 1: Simple (Step) Functions. If X is a **simple function** of the form

$$X(\omega) = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}(\omega),$$

where $\{A_i\}_{i=1}^n$ is a finite partition of Ω into measurable sets, then

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

Step 2: Non-negative Functions. If $X \geq 0$, the expectation is defined as the supremum over all simple functions s with $0 \leq s \leq X$:

$$\mathbb{E}[X] = \sup \{ \mathbb{E}[s] : s \text{ simple}, \ 0 \leq s \leq X \}.$$

Step 3: General Functions. Any measurable function X can be decomposed into its positive and negative parts:

$$X = X^+ - X^-,$$

where $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$. The expectation is then

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

provided that at least one of $\mathbb{E}[X^+]$ or $\mathbb{E}[X^-]$ is finite.

4.1.4 Existence and Integrability

A random variable X is said to be **integrable** (or to have **finite expectation**) if

$$\boxed{\mathbb{E}[|X|] = \int_{\Omega} |X(\omega)| \, d\mathbb{P}(\omega) < \infty}.$$

If $\mathbb{E}[|X|] = \infty$, the expectation is either undefined or infinite, as occurs in the **Cauchy distribution**, where both $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ diverge, yielding an indeterminacy of the form $\infty - \infty$.

4.1.5 Advantages over the Riemann Integral

The Lebesgue integral offers two key technical advantages:

1. **Geometric Unification:** While the Riemann integral partitions the **domain** of the function, the Lebesgue integral partitions the **codomain** (range), enabling integration of highly irregular functions and handling of sets of measure zero without complications.
2. **Superior Convergence Theorems:** The Lebesgue integral behaves exceptionally well under limits. The **Monotone Convergence Theorem** and the **Dominated Convergence Theorem** guarantee that, under general conditions, the limit of expectations equals the expectation of the limit—a property that fails for the Riemann integral in many cases.

4.2 Expectation of Functions of a Random Variable

4.2.1 Motivation

We frequently need to compute the expected value of a transformation $g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is a measurable function. The **Law of the Unconscious Statistician** (LOTUS) provides a direct method: we can compute $\mathbb{E}[g(X)]$ using the distribution of X alone, without first deriving the distribution of $Y = g(X)$.

4.2.2 Formal Statement (LOTUS)

For a measurable function $g : \mathbb{R} \to \mathbb{R}$,

$$\boxed{\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, d\mathbb{P}_X(x)}.$$

Case	Formula
Discrete X	$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\, p_X(x)$, provided the sum converges absolutely
Continuous X	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$, provided the integral converges absolutely

---

4.2.3 Existence Condition

The expectation $\mathbb{E}[g(X)]$ is well-defined only if

$$\mathbb{E}[|g(X)|] < \infty.$$

If this condition fails, the expectation may be infinite or undefined, as in the case of heavy-tailed distributions.

4.2.4 Fundamental Properties of the Expectation Operator

Property	Statement
Linearity	$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
Monotonicity	If $X \leq Y$ a.s., then $\mathbb{E}[X] \leq \mathbb{E}[Y]$
Positivity	If $X \geq 0$ a.s., then $\mathbb{E}[X] \geq 0$
Multiplicativity (Independence)	If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
Jensen's Inequality	If g is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$
Law of Iterated Expectations	$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$

4.2.5 Applications to Moments and Variance

The expectation operator enables the definition of distributional summaries:

- **Raw moments:** $\mu'_r = \mathbb{E}[X^r]$ for $r \in \mathbb{N}$.
- **Variance:** $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- **Standardization:** The standardized variable $Z = (X - \mu)/\sigma$ has $\mathbb{E}[Z] = 0$ and $\text{Var}(Z) = 1$.

4.3 Moments and the Moment Generating Function

4.3.1 Definition of Moments

Moments are expected values of powers of a random variable and serve as quantitative descriptors of its distributional shape.

4.3.1.1 Raw Moments (Moments about the Origin) For $r \in \mathbb{N}$, the **r-th raw moment** of X, denoted μ'_r, is

$$\boxed{\mu'_r = \mathbb{E}[X^r]}.$$

- **First raw moment** ($r = 1$): The population mean $\mu = \mathbb{E}[X]$.
- **Existence:** The r-th moment exists only if $\mathbb{E}[|X|^r] < \infty$.

4.3.1.2 Central Moments The **r-th central moment**, denoted μ_r, is defined relative to the mean:

$$\boxed{\mu_r = \mathbb{E}[(X - \mathbb{E}[X])^r]}.$$

Central Moment	Interpretation
$\mu_1 = \mathbb{E}[X - \mu]$	Always **zero**
$\mu_2 = \mathbb{E}[(X - \mu)^2]$	**Variance** (σ^2) — measures dispersion
$\mu_3 = \mathbb{E}[(X - \mu)^3]$	**Skewness** — measures asymmetry
$\mu_4 = \mathbb{E}[(X - \mu)^4]$	**Kurtosis** — measures tail weight / peakedness

Standardized Moments:

- **Skewness:** $\gamma_1 = \mu_3/\sigma^3$ (0 for symmetric distributions).
- **Excess Kurtosis:** $\gamma_2 = \mu_4/\sigma^4 - 3$ (0 for the normal distribution).

4.3.2 Moment Generating Function (MGF)

The **moment generating function** (MGF) is a technical tool that simplifies the computation of higher-order moments and facilitates proofs of convergence in distribution.

4.3.2.1 Formal Definition The MGF of a random variable X is defined as

$$\boxed{M_X(t) = \mathbb{E}[e^{tX}]}, \quad t \in \mathbb{R},$$

provided the expectation exists in a neighborhood of $t = 0$.

Case	Formula
Discrete	$M_X(t) = \sum_{x \in \mathcal{X}} e^{tx} p_X(x)$
Continuous	$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$

Existence Condition: The MGF is finite only if the sum or integral converges for t in an open interval containing 0.

4.3.2.2 Moment Generation Property The MGF "generates" the moments via differentiation at $t = 0$:

$$\boxed{\mathbb{E}[X^r] = \frac{d^r}{dt^r} M_X(t)\Big|_{t=0}}.$$

Example (Normal $N(\mu, \sigma^2)$):

- $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$
- $M'_X(0) = \mu$ (mean)
- $M''_X(0) = \mu^2 + \sigma^2$
- $\text{Var}(X) = M''_X(0) - [M'_X(0)]^2 = \sigma^2$

4.3.2.3 Key Properties of the MGF

Property	Statement
Uniqueness	If $M_X(t) = M_Y(t)$ for all t then $X \overset{d}{=} Y$
Linear Transformation	If $Y = aX + b$, then $M_Y(t$
Sum of Independent Variables	If $X \perp Y$, then $M_{X+Y}(t) =$
Continuity Theorem	If $M_{X_n}(t) \to M_X(t)$ pointw $X_n \xrightarrow{d} X$

4.3.2.4 MGFs of Common Distributions

Distribution	MGF $M_X(t)$	Dor
Bernoulli(p)	$1 - p + pe^t$	$\mathbb{R}$
Binomial(n, p)	$(1 - p + pe^t)^n$	$\mathbb{R}$
Poisson(λ)	$\exp\{\lambda(e^t - 1)\}$	$\mathbb{R}$
Normal(μ, σ^2)	$\exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$	$\mathbb{R}$
Exponential(λ)	$(1 - t/\lambda)^{-1}$	$t <$
Gamma(α, β)	$(1 - \beta t)^{-\alpha}$	$t <$
Chi-Square(ν)	$(1 - 2t)^{-\nu/2}$	$t <$
Uniform(a, b)	$\frac{e^{bt} - e^{at}}{t(b-a)}$	$\mathbb{R}$

4.3.2.5 Limitations and the Characteristic Function Not all distributions possess an MGF (e.g., the Cauchy distribution), because $\mathbb{E}[e^{tX}]$ may diverge for any $t > 0$. In such cases, the **characteristic function** is used:

$$\boxed{\phi_X(t) = \mathbb{E}[e^{itX}]}, \quad t \in \mathbb{R}.$$

The characteristic function **always exists** because $|e^{itX}| = 1$, ensuring boundedness and integrability. It shares the uniqueness and continuity properties of the MGF but is universally applicable.

4.4 Variance

4.4.1 Formal Definition

The **variance** of a random variable X, denoted $\mathrm{Var}(X)$ or σ^2, is defined as the expected squared deviation from its mean:

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

Case	Formula
Discrete	$\mathrm{Var}(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p_X(x)$
Continuous	$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\, dx$

4.4.2 Computational Identity

For practical computation, the following identity is often more convenient:

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

Proof:

1. $\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2]$.
2. By linearity, $= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2$.
3. Since $\mu = \mathbb{E}[X]$, $= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2$. $\square$

4.4.3 Properties of Variance

Property	Statement
Non-negativity	$\mathrm{Var}(X) \geq 0$
Constant	$\mathrm{Var}(c) = 0$ for any constant c
Linear Transformation	$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$
Sum (General)	$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$
Sum (Independent)	If $X \perp Y$, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
Linear Combination	$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) =$ $\sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$

4.4.4 Standard Deviation

Since variance is expressed in squared units of the original variable, the **standard deviation** σ is defined to restore the original units:

$$\boxed{\sigma = \sqrt{\mathrm{Var}(X)}}.$$

4.4.5 Sample Variance (S^2)

Given a sample $(X_1, \ldots, X_n)$, the sample variance is:

$$\boxed{S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2},$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

Unbiasedness: $\mathbb{E}[S^2] = \sigma^2$. The $n-1$ denominator (Bessel's correction) accounts for the loss of one degree of freedom due to estimating the mean.

4.5 Covariance

4.5.1 Formal Definition

The **covariance** between two random variables X and Y measures the degree of **linear association** between them:

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

Computational Identity:

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

Interpretation: - **Positive ($\mathrm{Cov} > 0$):** X and Y tend to move in the same direction. - **Negative ($\mathrm{Cov} < 0$):** X and Y tend to move in opposite directions. - **Zero ($\mathrm{Cov} = 0$):** No **linear** relationship.

4.5.2 Properties of Covariance

Property	Statement
Symmetry	$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
Self-Covariance	$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
Bilinearity	$\mathrm{Cov}(aX+b, cY+d) = ac \cdot \mathrm{Cov}(X,Y)$
Cauchy–Schwarz	$\lvert\mathrm{Cov}(X,Y)\rvert \le \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$

4.5.3 Covariance vs. Independence

The relationship between independence and covariance is subtle but crucial:

- **Independence implies zero covariance:** If $X \perp Y$, then $\mathrm{Cov}(X,Y) = 0$.
- **Zero covariance does NOT imply independence:** Variables can be nonlinearly dependent (e.g., $Y = X^2$ with $X \sim N(0,1)$) and still have zero covariance.
- **Exception — Bivariate Normal:** For jointly normal variables, zero covariance is equivalent to independence.

4.5.4 Correlation Coefficient

To obtain a scale-invariant measure, the **correlation coefficient** ρ is defined as:

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

Properties: - $-1 \le \rho \le 1$ (by Cauchy–Schwarz). - $\rho = 1$: perfect positive linear relationship. - $\rho = -1$: perfect negative linear relationship. - $\rho = 0$: no linear relationship.

4.5.5 Covariance Matrix (Σ)

For a random vector $\mathbf{X} = (X_1, \ldots, X_m)'$, the **covariance matrix** is the symmetric $m \times m$ matrix

$$\Sigma = \mathrm{Var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)'],$$

with elements

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j).$$

Properties: - Σ is symmetric and positive semidefinite. - The diagonal entries are the variances: $\Sigma_{ii} = \mathrm{Var}(X_i)$.

4.6 Important Inequalities

4.6.1 Markov's Inequality

Statement: If $X \geq 0$ almost surely and $\mathbb{E}[X] < \infty$, then for any $k > 0$,

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

General Form: For any $r > 0$ and $\epsilon > 0$,

$$\mathbb{P}(|X| \geq \epsilon) \leq \frac{\mathbb{E}[|X|^r]}{\epsilon^r}.$$

Proof (for $X \geq 0$):

1. Note that $X \geq k\mathbf{1}_{\{X \geq k\}}$.
2. Taking expectations: $\mathbb{E}[X] \geq k\mathbb{E}[\mathbf{1}_{\{X \geq k\}}] = k\mathbb{P}(X \geq k)$.
3. Therefore, $\mathbb{P}(X \geq k) \leq \mathbb{E}[X]/k$. $\square$

Role in Econometrics: Used to prove convergence in probability and the Weak Law of Large Numbers.

4.6.2 Chebyshev's Inequality

Statement: For any random variable X with finite variance, and for any $\epsilon > 0$,

$$\boxed{\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}}.$$

Alternative Form: For $k > 0$,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Example: $\mathbb{P}(|X - \mu| \geq 2\sigma) \leq 1/4 = 25\%$.

Proof: Apply Markov's inequality to $Y = (X - \mu)^2$ with $k = \epsilon^2$:

1. $\mathbb{P}((X - \mu)^2 \geq \epsilon^2) \leq \mathbb{E}[(X - \mu)^2]/\epsilon^2 = \mathrm{Var}(X)/\epsilon^2$.
2. Since $(X - \mu)^2 \geq \epsilon^2 \iff |X - \mu| \geq \epsilon$, the result follows. $\square$

Role in Econometrics: Used to prove the Weak Law of Large Numbers and the consistency of estimators.

4.6.3 Jensen's Inequality

Statement: Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathbb{E}[|X|] < \infty$, then

$$\boxed{\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])}.$$

If g is...	Inequality
Convex (e.g., $x^2, e^x, \lvert x\rvert$)	$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$
Concave (e.g., $\log x, \sqrt{x}$)	$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$

Equality occurs if: - g is linear ($g(x) = ax + b$), or - X is constant almost surely ($\mathrm{Var}(X) = 0$).

Examples: - $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ (non-negativity of variance). - $\mathbb{E}[\log X] \leq \log(\mathbb{E}[X])$ (for $X > 0$).

Conditional Version: For convex g,

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

Role in Econometrics: Used in MLE theory to prove that the true parameter maximizes the expected log-likelihood.

4.6.4 Cauchy–Schwarz Inequality

Probabilistic Form: For random variables X, Y with finite second moments,

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

Equivalently,

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2].$$

Vector Form: For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$,

$$|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|.$$

Integral Form (Lebesgue):

$$\int |fg|\, d\mu \leq \|f\|_{L^2}\|g\|_{L^2}.$$

Applications: - Proves that $|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$. - Hence, $|\rho| \leq 1$ and $0 \leq R^2 \leq 1$. - Used to prove the **Cramér–Rao Lower Bound**. - Used to prove that GLS is more efficient than OLS under heteroskedasticity. - Ensures that the product of two L^2 functions belongs to L^1.

4.7 Summary of Key Results

Concept	Formula	Key Condition
Expectation (General)	$\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P}$	$\mathbb{E}[\lvert X\rvert] < \infty$
Expectation (Discrete)	$\mathbb{E}[X] = \sum x_i p_X(x_i)$	Absolute convergence
Expectation (Continuous)	$\mathbb{E}[X] = \int x f_X(x)\, dx$	$\mathbb{E}[\lvert X\rvert] < \infty$
LOTUS	$\mathbb{E}[g(X)] = \int g(x)\, d\mathbb{P}_X(x)$	$\mathbb{E}[\lvert g(X)\rvert] < \infty$
Variance	$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$	$\mathbb{E}[X^2] < \infty$
Covariance	$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$	$\mathbb{E}[XY] < \infty$
MGF	$M_X(t) = \mathbb{E}[e^{tX}]$	Exists near $t = 0$
Markov	$\mathbb{P}(X \geq k) \leq \mathbb{E}[X]/k$	$X \geq 0$, $\mathbb{E}[X] < \infty$
Chebyshev	$\mathbb{P}(\lvert X-\mu\rvert \geq \epsilon) \leq \sigma^2/\epsilon^2$	$\sigma^2 < \infty$
Jensen	$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ if g convex	$\mathbb{E}[\lvert X\rvert] < \infty$
Cauchy–Schwarz	$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$	$\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$

Chapter 5: Random Vectors

5.1 Definition of Random Vectors

5.1.1 Motivation

A **random vector** (or multivariate random variable) extends the concept of a random variable to multiple dimensions, providing the mathematical framework for studying **joint distributions**. In econometrics, we rarely analyze a single variable in isolation—we study systems of simultaneous equations, portfolios of multiple assets, or vectors of individual characteristics. Random vectors enable the formal analysis of such multidimensional phenomena.

Geometric Interpretation in $\mathbb{R}^m$: Imagine each outcome $\omega \in \Omega$ as an abstract point. A random vector $\mathbf{X}$ maps each ω to a point in m-dimensional Euclidean space, where each component X_j corresponds to a coordinate. The distribution of $\mathbf{X}$ describes how probability mass is spread across this m-dimensional space.

5.1.2 Formal Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An m-dimensional **random vector** is a measurable function

$$\mathbf{X} : \Omega \to \mathbb{R}^m,$$

i.e., a function such that $\mathbf{X}^{-1}(B) \in \mathcal{F}$ for every Borel set $B \in \mathcal{B}(\mathbb{R}^m)$.

A random vector is composed of m individual random variables $X_1, X_2, \ldots, X_m$, all defined on the same probability space:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} = (X_1, X_2, \ldots, X_m)'.$$

Each component $X_i : \Omega \to \mathbb{R}$ is a real-valued random variable.

5.1.3 Joint Distribution Function

The probabilistic behavior of a random vector is fully described by its **joint cumulative distribution function (CDF)**:

$$\boxed{F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m)}.$$

Remark: The inequality $\mathbf{X} \leq \mathbf{x}$ is satisfied if and only if the condition holds for **all** components simultaneously.

5.1.4 Types of Random Vectors

Type	Description	Characterization
Discrete	Vector assumes values in a countable subset of $\mathbb{R}^m$	Joint probability mass function $p_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x})$
Continuous	There exists a non-negative integrable function $f_{\mathbf{X}}$	Joint probability density function (PDF)

For a continuous random vector, the joint PDF $f_{\mathbf{X}} : \mathbb{R}^m \to [0, \infty)$ satisfies

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}, \quad \forall A \in \mathcal{B}(\mathbb{R}^m),$$

with the normalization condition

$$\int_{\mathbb{R}^m} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = 1.$$

5.1.5 Moments of a Random Vector

For a random vector, the mean and variance take vector and matrix forms:

Moment	Definition
Mean Vector	$\mu = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_m])'$
Covariance Matrix	$\Sigma = \mathrm{Var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)']$

The covariance matrix is an $m \times m$ symmetric, positive semidefinite matrix with elements

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j).$$

Properties: - Diagonal entries: $\Sigma_{ii} = \mathrm{Var}(X_i)$. - Off-diagonal entries: $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \Sigma_{ji}$.

5.2 Joint Distributions

5.2.1 Joint CDF for Bivariate Case

For a pair of random variables (X, Y), the **joint CDF** is

$$\boxed{F(x, y) = \mathbb{P}(X \leq x, Y \leq y)}.$$

Properties:

1. **Bounds:** $0 \leq F(x, y) \leq 1$.

2. **Monotonicity:** F is non-decreasing in each argument.
3. **Limits:**

$$\lim_{x\to-\infty} F(x,y) = 0, \qquad \lim_{y\to-\infty} F(x,y) = 0,$$
$$\lim_{x\to\infty} F(x,y) = F_Y(y), \qquad \lim_{y\to\infty} F(x,y) = F_X(x),$$
$$\lim_{x,y\to\infty} F(x,y) = 1.$$

4. **Rectangle Probability:** For $a < b$ and $c < d$:

$$\mathbb{P}(a < X \le b, c < Y \le d) = F(b,d) - F(b,c) - F(a,d) + F(a,c).$$

5.2.2 Joint Probability Mass Function (Discrete)

If X and Y are discrete, the joint behavior is described by the **joint PMF**:

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

Properties:

1. $p(x,y) \ge 0$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$.
2. $\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) = 1$.

5.2.3 Joint Probability Density Function (Continuous)

For continuous variables, the **joint PDF** $f(x,y)$ is a surface over $\mathbb{R}^2$ such that the volume under the surface over a region R equals the probability:

$$\boxed{\mathbb{P}((X,Y) \in R) = \iint_R f(x,y)\,dx\,dy}.$$

Properties:

1. $f(x,y) \ge 0$ for all $(x,y) \in \mathbb{R}^2$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$.

Relationship with the Joint CDF: Where the joint CDF is sufficiently smooth,

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y).$$

5.3 Marginal Distributions

5.3.1 Motivation

The **marginal distribution** describes the individual behavior of one variable in a multivariate system, ignoring the values assumed by the other variables. While the joint distribution captures simultaneous behavior, the marginal distribution "summarizes" this information for a single dimension.

Geometric Interpretation: If the joint density is viewed as a surface in $\mathbb{R}^3$, the marginal density represents the projection or "shadow" of this surface onto one of the coordinate axes.

5.3.2 Continuous Case

For a pair of continuous random variables (X, Y) with joint PDF $f(x, y)$, the marginal densities are obtained by **integrating out** the variable we wish to remove:

Marginal	Formula
Marginal of X	$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$
Marginal of Y	$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$

5.3.3 Discrete Case

If the variables are discrete, the marginal probabilities are obtained by **summing** the joint PMF over all possible values of the other variable:

Marginal	Formula
Marginal of X	$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y)$
Marginal of Y	$p_Y(y) = \sum_{x \in \mathcal{X}} p(x, y)$

5.3.4 Properties of Marginal Distributions

1. **Validity:** Marginal densities are legitimate density functions:
   - Non-negative everywhere.
   - The integral (or sum) over the entire domain equals 1.
2. **Independence Criterion:** X and Y are independent if and only if
$$f(x, y) = f_X(x) f_Y(y) \quad \text{(continuous)},$$
$$p(x, y) = p_X(x) p_Y(y) \quad \text{(discrete)}.$$
3. **Marginal Expectation:** $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$.
4. **Relation to Conditional Densities:** The marginal density is the denominator in the definition of the conditional density:
$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}.$$

5.4 Conditional Distributions

5.4.1 Motivation

The **conditional distribution** describes the probability law of one random variable given that another has assumed a specific value. This concept is fundamental for understanding dependence relationships and serves as the basis for regression analysis and conditional expectation.

Geometric Interpretation: The conditional density can be visualized as a **"slice" of the joint density surface** along a fixed value of x, which is then **renormalized** (divided by $f_X(x)$) so that the area under this slice equals 1.

5.4.2 Formal Definition (Continuous Case)

For continuous random variables (X, Y) with joint PDF $f(x, y)$ and marginal density $f_X(x) > 0$, the conditional density of Y given $X = x$ is

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

Similarly,

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

5.4.3 Discrete Case

For discrete variables,

$$\boxed{\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}}, \quad \mathbb{P}(X = x) > 0.$$

5.4.4 Properties of Conditional Densities

1. **Validity:** For each fixed x, $f_{Y|X}(y \mid x)$ is a legitimate density:
   - Non-negative.
   - $\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\, dy = 1$.
2. **Independence:** If X and Y are independent, then
$$f_{Y|X}(y \mid x) = f_Y(y), \qquad f_{X|Y}(x \mid y) = f_X(x).$$
3. **Bayes' Theorem for Densities:**
$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y)\, f_Y(y)}{f_X(x)}.$$
4. **Conditional Expectation:**
$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy.$$
5. **Conditional Variance:**
$$\mathrm{Var}(Y \mid X = x) = \int_{-\infty}^{\infty} (y - \mathbb{E}[Y \mid X = x])^2 f_{Y|X}(y \mid x)\, dy.$$
6. **Bivariate Normal:** In the bivariate normal model, the conditional densities $f_{Y|X}$ and $f_{X|Y}$ are always **univariate normal distributions**.

5.4.5 Law of Total Expectation (LTE)

The **Law of Total Expectation** (or Law of Iterated Expectations) states:

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

Continuous Case:

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

Discrete Case:

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

5.4.6 Law of Total Variance (LTV)

The **Law of Total Variance** decomposes the total variance into two components:

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

Interpretation: - $\mathbb{E}[\mathrm{Var}(Y \mid X)]$: Average within-group variance (unexplained variation). - $\mathrm{Var}(\mathbb{E}[Y \mid X])$: Variance of the conditional means (explained variation).

5.5 Independence of Random Variables

5.5.1 Motivation

Independence describes situations where the outcome of one variable provides no information about the outcome of another. Two random variables X and Y are statistically independent if the occurrence of any event associated with X does not alter the probability of events associated with Y.

5.5.2 Formal Definitions

Independence can be characterized through various equivalent conditions:

Criterion	Condition
CDF	$F(x, y) = F_X(x)F_Y(y)$ for all x, y
PDF (Continuous)	$f(x, y) = f_X(x)f_Y(y)$ for all x, y
PMF (Discrete)	$p(x, y) = p_X(x)p_Y(y)$ for all x, y
Conditional	$f_{Y\|X}(y \mid x) = f_Y(y)$ and $f_{X\|Y}(x \mid y) = f_X(x)$

5.5.3 Properties of Independence

Property	Statement
Expectation of Product	$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
General Functions	$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$
Transformation	If $X \perp Y$, then $g(X) \perp h(Y)$ for any measurable g, h
Variance of Sum	$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
Conditional Expectation	$\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$
Conditional Variance	$\mathrm{Var}(Y \mid X) = \mathrm{Var}(Y)$

5.5.4 Independence vs. Zero Correlation

A crucial distinction:

Relationship	Statement
Independence $\implies$ Zero Correlation	If $X \perp Y$, then $\text{Cov}(X, Y) = 0$ and $\rho = 0$
Zero Correlation $\not\Rightarrow$ Independence	Variables can be nonlinearly dependent (e.g., $Y = X^2$ with $X \sim N(0, 1)$) yet have zero correlation
Bivariate Normal Exception	If (X, Y) is jointly normal, zero correlation ($\rho = 0$) is equivalent to independence

---

5.5.5 Mutual Independence vs. Pairwise Independence

- **Mutual Independence:** A set of variables $\{X_1, \ldots, X_n\}$ is mutually independent if the joint density of any subset factors as the product of the marginals:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

- **Pairwise Independence:** Independence between each pair—i.e., $X_i \perp X_j$ for all $i \neq j$—does **not** imply mutual independence of the entire set.
- **i.i.d. (Independent and Identically Distributed):** A sequence $\{X_i\}_{i=1}^n$ is i.i.d. if the variables are mutually independent and each has the same marginal distribution.

5.6 Transformations of Random Vectors

5.6.1 Motivation

The transformation of a random vector through a **bijective** (one-to-one) and differentiable function is a fundamental technique for deriving new probability distributions from known ones. This process uses the **Jacobian determinant** to adjust the probability density in the transformed space.

Geometric Interpretation: The Jacobian determinant compensates for how the transformation "stretches" or "shrinks" areas (in 2D) or volumes (in higher dimensions) during the change of coordinates.

5.6.2 The Change-of-Variables Theorem

Let $\mathbf{X}$ be an m-dimensional random vector with joint PDF $f_{\mathbf{X}}(\mathbf{x})$. Consider a transformation $\mathbf{Y} = g(\mathbf{X})$, where $g : \mathbb{R}^m \to \mathbb{R}^m$ is a **one-to-one** differentiable function. The joint PDF of $\mathbf{Y}$ is

$$\boxed{f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y})) \cdot |\det(J(\mathbf{y}))|},$$

where: - $h(\mathbf{y}) = g^{-1}(\mathbf{y})$ is the inverse function. - $J(\mathbf{y})$ is the **Jacobian matrix** of the inverse transformation. - $|\det(J(\mathbf{y}))|$ is the absolute value of the determinant.

5.6.3 The Jacobian Matrix

For an inverse transformation $h(\mathbf{y}) = (h_1(\mathbf{y}), \ldots, h_m(\mathbf{y}))'$, the Jacobian matrix is

$$J(\mathbf{y}) = \begin{pmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} & \cdots & \frac{\partial h_1}{\partial y_m} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} & \cdots & \frac{\partial h_2}{\partial y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_m}{\partial y_1} & \frac{\partial h_m}{\partial y_2} & \cdots & \frac{\partial h_m}{\partial y_m} \end{pmatrix}.$$

Requirements:

1. **Bijectivity:** The transformation must be one-to-one on the support of $\mathbf{X}$.
2. **Differentiability:** Both g and its inverse h must have continuous partial derivatives.
3. **Non-zero Jacobian:** $\det(J(\mathbf{y})) \neq 0$ on the support of the distribution.

5.6.4 Special Cases

5.6.4.1 Univariate Case For $Y = g(X)$ with g strictly monotone and differentiable,

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|.$$

5.6.4.2 Linear Transformation If $\mathbf{Y} = A\mathbf{X}$, where A is an invertible $m \times m$ matrix, then

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(A^{-1}\mathbf{y}) \cdot |\det(A)|^{-1}.$$

5.6.4.3 Non-Monotone Transformations If $g(x)$ is not monotone (e.g., $Y = X^2$), the density of Y is obtained by summing over the different branches:

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left[f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right], \quad y > 0.$$

5.6.5 Example: Sum of Exponentials

Let X_1, X_2 be independent with densities $f_{X_1}(x_1) = e^{-x_1}$ and $f_{X_2}(x_2) = e^{-x_2}$ for $x_1, x_2 \geq 0$. Define

$$Y_1 = X_1, \qquad Y_2 = X_1 + X_2.$$

The inverse transformation is

$$X_1 = Y_1, \qquad X_2 = Y_2 - Y_1,$$

with support $\{0 \leq y_1 \leq y_2 < \infty\}$. The Jacobian matrix is

$$J = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \quad \det(J) = 1.$$

The joint density of (Y_1, Y_2) is

$$f_{Y_1,Y_2}(y_1, y_2) = e^{-y_1} e^{-(y_2 - y_1)} = e^{-y_2}, \quad 0 \leq y_1 \leq y_2.$$

The marginal density of Y_2 (the sum) is

$$f_{Y_2}(y_2) = \int_0^{y_2} e^{-y_2} \, dy_1 = y_2 e^{-y_2},$$

which is the Gamma$(2, 1)$ distribution.

5.6.6 Application: Box–Muller Transformation

The Box–Muller transformation is a classic application of the change-of-variables theorem. It transforms a pair of independent uniform variables $(U_1, U_2) \sim \text{Unif}(0, 1)^2$ into a pair of independent standard normal variables:

$$Z_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2), \qquad Z_2 = \sqrt{-2 \log U_1} \sin(2\pi U_2).$$

The Jacobian of this transformation yields the product of two standard normal densities.

5.7 Summary Table

Concept	Formula
Joint CDF	$F(x, y) = \mathbb{P}(X \leq x, Y \leq y)$
Joint PDF (Continuous)	$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$
Joint PMF (Discrete)	$p(x, y) = \mathbb{P}(X = x, Y = y)$
Marginal PDF of X	$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$
Marginal PMF of X	$p_X(x) = \sum_y p(x, y)$
Conditional PDF	$f_{Y \mid X}(y \mid x) = \frac{f(x,y)}{f_X(x)}$
Conditional PMF	$\mathbb{P}(Y = y \mid X = x) = \frac{p(x,y)}{p_X(x)}$
Independence	$f(x, y) = f_X(x) f_Y(y)$ or $p(x, y) = p_X(x) p_Y(y)$
Law of Total Expectation	$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$
Law of Total Variance	$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y \mid X)] + \text{Var}(\mathbb{E}[Y \mid X])$
Change of Variables	$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(h(\mathbf{y})) \cdot \lvert \det(J(\mathbf{y})) \rvert$
Mean Vector	$\mu = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_m])'$
Covariance Matrix	$\Sigma = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)']$

Chapter 6: Conditional Expectation

6.1 Conditional Expectation Given a σ-Algebra

6.1.1 Motivation

The **conditional expectation given a σ-algebra** provides the most rigorous and general formalization of conditional expectation within measure theory. While elementary treatments focus on conditioning on a specific event or random variable, the σ-algebra approach enables conditioning on an entire **information set**—a collection of events whose occurrence we can observe. This framework is essential for stochastic processes, martingale theory, and advanced econometrics.

Geometric Interpretation: Think of $\mathcal{G}$ as a "filter" or "lens" that only reveals certain events. The conditional expectation $\mathbb{E}[Y \mid \mathcal{G}]$ is the **best approximation** of Y using only information available through this filter. In L^2 terms, it is the orthogonal projection of Y onto the subspace of $\mathcal{G}$-measurable functions.

6.1.2 Formal Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let Y be a random variable with finite expectation, i.e., $\mathbb{E}[|Y|] < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$.

The **conditional expectation of Y given $\mathcal{G}$**, denoted $\mathbb{E}[Y \mid \mathcal{G}]$, is a random variable $Z : \Omega \to \mathbb{R}$ that satisfies two fundamental conditions:

1. **Measurability:** $Z = \mathbb{E}[Y \mid \mathcal{G}]$ is **$\mathcal{G}$-measurable**. This means the value of the conditional expectation depends only on the information contained in $\mathcal{G}$.
2. **Integral Identity:** For every $A \in \mathcal{G}$,

$$\boxed{\int_A \mathbb{E}[Y \mid \mathcal{G}]\, d\mathbb{P} = \int_A Y\, d\mathbb{P}}.$$

Equivalently, for all $\mathcal{G}$-measurable bounded functions g,

$$\mathbb{E}[g \cdot \mathbb{E}[Y \mid \mathcal{G}]] = \mathbb{E}[gY].$$

6.1.3 Existence and Uniqueness

The existence of $\mathbb{E}[Y \mid \mathcal{G}]$ is guaranteed by the **Radon–Nikodým theorem**. Specifically, define a signed measure ν on $(\Omega, \mathcal{G})$ by

$$\nu(A) = \int_A Y\, d\mathbb{P}, \quad A \in \mathcal{G}.$$

Since $\nu \ll \mathbb{P}$ on $\mathcal{G}$, the Radon–Nikodým theorem ensures the existence of a $\mathcal{G}$-measurable function Z such that $\nu(A) = \int_A Z\, d\mathbb{P}$ for all $A \in \mathcal{G}$. This Z is precisely $\mathbb{E}[Y \mid \mathcal{G}]$.

Uniqueness: The function Z is unique **almost surely**—any two versions differ only on a set of probability zero.

6.1.4 Interpretation as an Information Set

In econometrics and time series analysis, the σ-algebra $\mathcal{G}$ is often interpreted as an **information set** (e.g., the past history of a series up to time t).

$\mathcal{G}$	Interpretation	Result
$\mathcal{G} = \{\emptyset, \Omega\}$	Trivial σ-algebra (no information)	$\mathbb{E}[Y \mid \mathcal{G}] = \mathbb{E}[Y]$
$\mathcal{G} = \mathcal{F}$	Full information	$\mathbb{E}[Y \mid \mathcal{G}] = Y$ almost surely
$\mathcal{G} = \mathcal{F}_{t-1}$	Information up to time $t-1$	Forms the basis for martingale difference sequences (MDS)

6.1.5 Relation to $\mathbb{E}[Y \mid X]$

The conditional expectation given a random variable X is a special case where the σ-algebra is the one **generated by** X:

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

Here, $\sigma(X)$ represents the collection of all events whose occurrence can be determined by knowing the value of X. By the Doob–Dynkin lemma, $\mathbb{E}[Y \mid X]$ is always a measurable function of X: there exists $m : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[Y \mid X] = m(X)$.

6.1.6 The Law of Iterated Expectations (Generalized Form)

The **Law of Iterated Expectations** (LIE), also known as the **tower property**, is a fundamental property. If $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ (i.e., $\mathcal{G}_1$ contains less information than $\mathcal{G}_2$), then

$$\boxed{\mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{G}_2] \mid \mathcal{G}_1\big] = \mathbb{E}[Y \mid \mathcal{G}_1]}.$$

Key Principle: The smallest information set prevails. When successively projecting a variable onto different levels of information, the final result is always determined by the **least refined information set** in the sequence.

Proof Sketch: For any $A \in \mathcal{G}_1 \subseteq \mathcal{G}_2$,

$$\int_A \mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}_2] \mid \mathcal{G}_1]\, d\mathbb{P} = \int_A \mathbb{E}[Y \mid \mathcal{G}_2]\, d\mathbb{P} = \int_A Y\, d\mathbb{P}.$$

The result follows by the uniqueness of conditional expectation. □

6.2 Conditional Expectation Given a Random Variable

6.2.1 Motivation

The **conditional expectation of Y given X**, often denoted $\mathbb{E}[Y \mid X]$ and called the **conditional expectation function** (CEF), describes the average value of Y when X takes a specific value x. It represents the central tendency of the distribution of Y within the subpopulation where $X = x$. This is the fundamental object of regression analysis.

6.2.2 Formal Definitions

The computation depends on the nature of the variables involved.

6.2.2.1 Discrete Case If X and Y are discrete,

$$\boxed{\mathbb{E}[Y \mid X = x] = \sum_{y \in \mathcal{Y}} y\, \mathbb{P}(Y = y \mid X = x)},$$

where

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}, \quad \mathbb{P}(X = x) > 0.$$

6.2.2.2 Continuous Case If (X, Y) has joint density $f(x, y)$,

$$\boxed{\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy = \frac{\int_{-\infty}^{\infty} y\, f(x, y)\, dy}{f_X(x)}},$$

for any x such that $f_X(x) > 0$.

6.2.2.3 General (Measure-Theoretic) Case For arbitrary random variables,

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

6.2.3 The Distinction: $m(x)$ vs. $m(X)$

A crucial distinction in practice:

Notation	Type	Description
$\mathbb{E}[Y \mid X = x]$ or $m(x)$	**Number (constant)**	The mean of Y for a specific observed value $X = x$
$\mathbb{E}[Y \mid X]$ or $m(X)$	**Random Variable**	A function of the random variable X; its value varies across the sample

6.2.4 Properties of Conditional Expectation

Property	Statement
Linearity	$\mathbb{E}[aY + bZ \mid X] = a\mathbb{E}[Y \mid X] + b\mathbb{E}[Z \mid X]$
Law of Iterated Expectations	$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$
Conditioning Theorem (Substitution)	$\mathbb{E}[g(X)Y \mid X] = g(X)\mathbb{E}[Y \mid X]$
Self-Conditioning	$\mathbb{E}[X \mid X] = X$
Function of X	$\mathbb{E}[g(X) \mid X] = g(X)$
Independence	If $X \perp Y$, then $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$
Tower Property	If $\sigma(X_1) \subseteq \sigma(X_2)$, then $\mathbb{E}[\mathbb{E}[Y \mid X_2] \mid X_1] = \mathbb{E}[Y \mid X_1]$
Monotonicity	If $Y \leq Z$ a.s., then $\mathbb{E}[Y \mid X] \leq \mathbb{E}[Z \mid X]$
Jensen (Conditional)	If g is convex, then $g(\mathbb{E}[Y \mid X]) \leq \mathbb{E}[g(Y) \mid X]$

6.2.5 Conditional Expectation as the Best Predictor

A central result in decision theory is that the conditional expectation $m(X) = \mathbb{E}[Y \mid X]$ is the **best predictor of** Y given X in the Mean Squared Error (MSE) sense:

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

for **any** measurable function $g : \mathbb{R} \to \mathbb{R}$. Equality holds if and only if $g(X) = \mathbb{E}[Y \mid X]$ almost surely.

Proof Sketch: For any $g(X)$,

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

because the cross-term vanishes by the conditioning theorem. □

6.2.6 The Regression Error and Exogeneity

Define the **regression error** as

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

By construction, the error has **zero conditional mean**:

$$\boxed{\mathbb{E}[e \mid X] = 0}.$$

This implies: - $\mathbb{E}[e] = 0$ (unconditional zero mean). - $\mathrm{Cov}(e, g(X)) = 0$ for any measurable function g. - $\mathbb{E}[e \cdot g(X)] = 0$ (orthogonality condition).

The error e is **uncorrelated with any function of** X, making it the foundation for the exogeneity assumption in regression models.

6.2.7 Special Cases

6.2.7.1 Linear Regression Model In the model $Y = \beta_0 + \beta_1 X + u$ with $\mathbb{E}[u \mid X] = 0$,

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X.$$

6.2.7.2 Bivariate Normal If (X, Y) follows a bivariate normal distribution,

$$\mathbb{E}[Y \mid X = x] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

where $\rho = \text{Corr}(X, Y)$ is the correlation coefficient.

6.3 Conditioning and Measurable Variables

6.3.1 The Doob–Dynkin Factorization Lemma

The **Doob–Dynkin Lemma** establishes an equivalence between informational dependence (via σ-algebras) and functional dependence.

Theorem: A random variable Y is measurable with respect to the σ-algebra generated by X, denoted $\sigma(X)$, **if and only if** there exists a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that

$$\boxed{Y = f(X)}.$$

6.3.2 Implications

Direction	Implication
Forward	If $Y = f(X)$ with f measurable, then Y is $\sigma(X)$-measurable
Inverse (Factorization)	If Y is $\sigma(X)$-measurable, then all information needed to determine Y is contained in X; hence $Y = f(X)$ for some measurable f

---

6.3.3 Application to Conditional Expectation

This lemma justifies why $\mathbb{E}[Y \mid X]$ is always treated as a function of X:

- By definition, $\mathbb{E}[Y \mid X]$ is measurable with respect to $\sigma(X)$.
- Therefore, by the Doob–Dynkin lemma, there exists a function $m(x)$ such that

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

This function $m(\cdot)$ is the **conditional expectation function (CEF)**.

6.3.4 The Conditioning Theorem (Substitution)

If $g(X)$ is a measurable function of X, it is automatically $\sigma(X)$-measurable. This allows us to "pull it out" of the conditional expectation operator:

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

This property is vital for proving exogeneity of error terms and orthogonality conditions in regression models.

6.4 Conditional Expectation and Independence

6.4.1 Independence as a Sufficient Condition

If X and Y are **statistically independent**, knowledge of X provides no information about Y. Mathematically,

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

The best prediction of Y given X is simply its unconditional mean.

6.4.2 Mean Independence

In econometrics, we often do not require full independence, but rather **mean independence**:

Concept	Definition
Mean Independence	$\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$
Full Independence	$f(x, y) = f_X(x) f_Y(y)$

Implication: Full independence $\implies$ Mean independence, but the **converse is not true**.

Example: Let $e = Xu$, where X and u are independent with $u \sim N(0, 1)$ and $X \sim N(0, 1)$. Then: - $\mathbb{E}[e \mid X] = X\mathbb{E}[u] = 0$ (mean independence). - But $\text{Var}(e \mid X) = X^2\text{Var}(u) = X^2$ depends on X, so e and X are not fully independent.

6.4.3 Conditional Independence Assumption (CIA)

A more advanced concept, central to causal identification, is the **Conditional Independence Assumption** (CIA), also known as *unconfoundedness*:

$$\boxed{Y_0, Y_1 \perp D \mid X}.$$

After controlling for observables X, the treatment indicator D becomes independent of the potential outcomes (Y_0, Y_1). This allows treatment assignment to be treated as "as good as random" within strata defined by X.

6.4.4 Independence vs. Covariance

Relationship	Statement
Independence $\implies$ Zero Covariance	If $X \perp Y$, then $\text{Cov}(X, Y) = 0$
Zero Covariance $\not\Rightarrow$ Independence	Variables can have zero covariance without being independent
Bivariate Normal Exception	For jointly normal variables, zero covariance $\iff$ independence

---

6.5 Conditional Variance

6.5.1 Formal Definition

Let Y be a random variable with finite second moment ($\mathbb{E}[Y^2] < \infty$). The **conditional variance** of Y given $X = x$ is

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

As a random variable:

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

Unlike the unconditional variance $\mathrm{Var}(Y)$ (a constant), the conditional variance is a **function of the regressors** X.

6.5.2 Computational Identity

As with ordinary variance,

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

6.5.3 Properties

Property	Statement
Non-negativity	$\mathrm{Var}(Y \mid X) \geq 0$ almost surely
Independence	If $X \perp Y$, then $\mathrm{Var}(Y \mid X) = \mathrm{Var}(Y)$
Linear Transformation	$\mathrm{Var}[a(X)Y + b(X) \mid X] = [a(X)]^2\mathrm{Var}[Y \mid X]$
Error Variance	If $e = Y - \mathbb{E}[Y \mid X]$, then $\mathrm{Var}(e \mid X) = \mathrm{Var}(Y \mid X) = \sigma^2(X)$

6.5.4 Law of Total Variance (Variance Decomposition)

One of the most important results is the identity relating total variance to its conditional components:

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

Component	Interpretation
$\mathbb{E}[\mathrm{Var}(Y \mid X)]$	**Within-group variance** — variation of Y **not explained** by X
$\mathrm{Var}(\mathbb{E}[Y \mid X])$	**Between-group variance** — variation of Y **explained** by X

Proof: Let $\mu(X) = \mathbb{E}[Y \mid X]$. Then

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

Expanding and using the fact that the cross-term is zero (by the conditioning theorem) yields the result. □

6.5.5 Applications: Homoskedasticity vs. Heteroskedasticity

These fundamental econometric concepts refer directly to the nature of the conditional variance:

Term	Condition
Homoskedasticity	$\sigma^2(x) = \sigma^2$ (constant, independent of x)
Heteroskedasticity	$\sigma^2(x)$ **varies** with x (e.g., consumption variance increases with income)

6.5.6 Geometric Interpretation and R^2

In linear regression models, the Law of Total Variance corresponds to the **decomposition of the sum of squares**:

Component	Corresponds to
Total Sum of Squares (SST)	Total variation of Y: $\sum(Y_i - \bar{Y})^2$
Explained Sum of Squares (SSE)	Variation explained by the model: $\sum(\hat{Y}_i - \bar{Y})^2$
Residual Sum of Squares (SSR)	Unexplained variation: $\sum(Y_i - \hat{Y}_i)^2$

The decomposition follows the **Pythagorean theorem**, since fitted values and residuals are orthogonal in OLS:

$$SST = SSE + SSR.$$

The coefficient of determination is

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

In population terms, for the CEF,

$$R^2 = \frac{\mathrm{Var}(\mathbb{E}[Y \mid X])}{\mathrm{Var}(Y)} = 1 - \frac{\mathbb{E}[\mathrm{Var}(Y \mid X)]}{\mathrm{Var}(Y)}.$$

6.6 Summary of Key Results

Concept	Formula
CE given σ-algebra	$\mathbb{E}[Y \mid \mathcal{G}]$ is $\mathcal{G}$-measurable; $\int_A \mathbb{E}[Y \mid \mathcal{G}]d\mathbb{P} = \int_A Y d\mathbb{P}$ for all $A \in \mathcal{G}$
CE given X	$\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid \sigma(X)]$
Law of Iterated Expectations	$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$
Generalized LIE	$\mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}_2] \mid \mathcal{G}_1] = \mathbb{E}[Y \mid \mathcal{G}_1]$ for $\mathcal{G}_1 \subseteq \mathcal{G}_2$
Conditioning Theorem	$\mathbb{E}[g(X)Y \mid X] = g(X)\mathbb{E}[Y \mid X]$
Best Predictor	$\mathbb{E}[Y \mid X]$ minimizes MSE among all functions of X
Mean Independence	$\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$
Conditional Variance	$\mathrm{Var}(Y \mid X) = \mathbb{E}[Y^2 \mid X] - (\mathbb{E}[Y \mid X])^2$
Law of Total Variance	$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])$
Regression Error	$e = Y - \mathbb{E}[Y \mid X]$ has $\mathbb{E}[e \mid X] = 0$
Doob–Dynkin Lemma	If Y is $\sigma(X)$-measurable, then $Y = f(X)$ for some measurable f

Chapter 7: Convergence of Random Variables

7.1 Convergence in Probability

7.1.1 Motivation

Convergence in probability is a fundamental concept in asymptotic theory, describing how a sequence of random variables concentrates around a target value as the sample size increases. It provides the mathematical foundation for the **consistency** of estimators in econometrics—the property that an estimator approaches the true parameter value as more data become available.

Geometric Interpretation: Imagine throwing darts at a bullseye. Convergence in probability means that as the number of throws increases, the probability that a dart lands far from the bullseye becomes arbitrarily small. The darts may occasionally land far away, but such occurrences become increasingly rare.

7.1.2 Formal Definition

A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in probability** to a random variable X (which may be a constant c) if, for every $\epsilon > 0$,

$$\boxed{\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0}.$$

Equivalently,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1.$$

Notation: $X_n \xrightarrow{p} X$ or $\text{plim}_{n\to\infty} X_n = X$, where **plim** denotes the *probability limit*.

Remark: If the limit is a constant c, then $\text{plim}(c) = c$.

7.1.3 Extension to Vectors and Matrices

A random vector $\mathbf{X}_n \in \mathbb{R}^k$ converges in probability to $\mathbf{X}$ if and only if **each component** converges in probability to the corresponding component of the limit:

$$\mathbf{X}_n \xrightarrow{p} \mathbf{X} \iff X_{n,j} \xrightarrow{p} X_j \quad \text{for all } j = 1, \ldots, k.$$

The same property holds for random matrices by vectorization.

7.1.4 Properties of the Probability Limit

The **plim** operator behaves analogously to deterministic limits under algebraic operations:

Operation	Formula
Sum/Difference	$\text{plim}(X_n \pm Y_n) = \text{plim}\, X_n \pm \text{plim}\, Y_n$
Product	$\text{plim}(X_n Y_n) = (\text{plim}\, X_n)(\text{plim}\, Y_n)$
Division	$\text{plim}(X_n/Y_n) = (\text{plim}\, X_n)/(\text{plim}\, Y_n)$, provided $\text{plim}\, Y_n \neq 0$
Inverse (Matrix)	If $A_n \xrightarrow{p} A$ with A non-singular, then $A_n^{-1} \xrightarrow{p} A^{-1}$

7.1.5 Continuous Mapping Theorem

If $X_n \xrightarrow{p} c$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous at c, then

$$\boxed{g(X_n) \xrightarrow{p} g(c)}.$$

Implication: If $\hat{\theta}_n$ is a consistent estimator of θ, then $g(\hat{\theta}_n)$ is consistent for $g(\theta)$ for any continuous function g. For example, if $\hat{\theta} \xrightarrow{p} \theta$, then $\log(\hat{\theta}) \xrightarrow{p} \log(\theta)$ (provided $\theta > 0$) and $1/\hat{\theta} \xrightarrow{p} 1/\theta$ (provided $\theta \neq 0$).

7.1.6 Relationship with Other Modes

Convergence in probability occupies an intermediate position in the hierarchy:

Implication	Direction
Almost Sure Convergence ($\xrightarrow{a.s.}$)	$\implies$ Convergence in Probability
Mean Square Convergence (L_2)	$\implies$ Convergence in Probability
Convergence in Probability	$\implies$ Convergence in Distribution ($\xrightarrow{d}$)
Constant Limit Equivalence	$X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$

7.1.7 Main Applications

1. **Weak Law of Large Numbers (WLLN):** The sample mean $\bar{X}_n$ converges in probability to the population mean μ.
2. **Consistency of Estimators:** An estimator $\hat{\theta}_n$ is **consistent** if $\hat{\theta}_n \xrightarrow{p} \theta$. A sufficient condition for consistency is that both the bias and the variance of the estimator tend to zero as $n \to \infty$:

$$\mathbb{E}[\hat{\theta}_n] \to \theta \quad \text{and} \quad \text{Var}(\hat{\theta}_n) \to 0.$$

7.2 Almost Sure Convergence

7.2.1 Motivation

Almost sure convergence (or *strong convergence*) is one of the most rigorous modes of convergence in probability theory. It describes a situation where the sequence of random variables behaves, in the limit, like a deterministic sequence for almost every possible outcome.

Geometric Interpretation: Imagine observing an infinite sequence of coin tosses. Almost sure convergence means that for *every possible infinite sequence of outcomes*—except perhaps a set of measure zero—the cumulative proportion of heads converges to 1/2.

7.2.2 Formal Definition

A sequence of random variables $\{Z_n\}_{n=1}^{\infty}$ converges **almost surely** to a constant c (or to a random variable Z) if

$$\boxed{\mathbb{P}\left(\lim_{n\to\infty} Z_n = c\right) = 1}.$$

Equivalently, for every $\epsilon > 0$,

$$\mathbb{P}\left(\limsup_{n\to\infty} |Z_n - c| > \epsilon\right) = 0.$$

This means that the set of outcomes for which the sequence fails to converge to the limit has probability zero.

Notation: $Z_n \xrightarrow{a.s.} c$ or $Z_n \to c$ almost surely.

7.2.3 Almost Sure vs. Convergence in Probability

Convergence in Probability	Almost Sure Convergence
For sufficiently large n, the probability that Z_n is "far" from the limit is small	The entire trajectory of the sequence approaches and **stays** close to the limit as $n \to \infty$
Allows occasional large deviations as long as they become rare	Requires that large deviations eventually cease to occur altogether (with probability 1)
Weaker mode	Stronger mode

Counterexample: A sequence can converge in probability but not almost surely. For example, consider a sequence that occasionally takes a large value with decreasing probability; the probability of seeing a large value at any fixed n tends to zero, but the sequence may still take large values infinitely often.

7.2.4 Relationship Between Modes

Almost sure convergence sits at the top of the hierarchy:

$$\boxed{\text{Almost Sure} \implies \text{Probability} \implies \text{Distribution}},$$

and

$$\boxed{\text{Mean Square } (L_2) \implies \text{Probability} \implies \text{Distribution}}.$$

The converses are not true in general.

7.2.5 The Borel–Cantelli Lemma

The **Borel–Cantelli Lemma** is the fundamental tool for establishing almost sure convergence.

First Borel–Cantelli Lemma: If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then

$$\mathbb{P}(A_n \text{ occurs infinitely often}) = 0.$$

Second Borel–Cantelli Lemma: If $\{A_n\}$ are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then

$$\mathbb{P}(A_n \text{ occurs infinitely often}) = 1.$$

Application: To prove $X_n \xrightarrow{a.s.} X$, it suffices to show that for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \epsilon) < \infty.$$

7.2.6 Strong Law of Large Numbers (SLLN)

The main application of almost sure convergence is the **Strong Law of Large Numbers**. If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. with finite expectation $\mathbb{E}[|X|] < \infty$, then

$$\boxed{\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu = \mathbb{E}[X]}.$$

Key Difference from WLLN: The Strong Law guarantees that the sample mean from a **single infinitely growing sample** will converge to the true value with "practical certainty" (probability 1). The Weak Law only guarantees that the probability of deviation becomes small.

7.2.7 Strong Consistency

An estimator $\hat{\theta}_n$ is **strongly consistent** if

$$\hat{\theta}_n \xrightarrow{a.s.} \theta.$$

Strong consistency is a more demanding property than ordinary (weak) consistency, but it provides stronger guarantees for the behavior of estimators.

7.3 Convergence in Distribution

7.3.1 Motivation

Convergence in distribution (also called *weak convergence* or *convergence in law*) describes how the probabilistic behavior of a sequence of random variables approaches a specific distribution as the sample size increases. Unlike convergence in probability, which focuses on the proximity of *values*, convergence in distribution focuses on the **shape of the distribution**.

Geometric Interpretation: The CDFs of the sequence become increasingly similar to the CDF of the limiting distribution.

7.3.2 Formal Definition

A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ with CDFs $F_n(x)$ converges **in distribution** to a random variable X with CDF $F(x)$ if

$$\boxed{\lim_{n\to\infty} F_n(x) = F(x)}$$

for all points $x \in \mathbb{R}$ where F is **continuous**.

Notation: $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{L}} X$.

Remark: The condition is required only at continuity points of F because the CDF may have jumps, and convergence at discontinuity points is not guaranteed.

7.3.3 Relationship with Other Modes

Relationship	Statement
Implication	Convergence in probability $\implies$ Convergence in distribution
Converse	Not true in general (convergence in distribution does not imply convergence in probability)
Constant Limit Equivalence	$X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$

7.3.4 Multivariate Convergence and the Cramér–Wold Device

For random vectors $\mathbf{X}_n \in \mathbb{R}^k$, convergence in distribution requires that the joint CDF of $\mathbf{X}_n$ converges to the joint CDF of $\mathbf{X}$. The **Cramér–Wold Device** simplifies this to checking all linear combinations:

$$\boxed{\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \lambda'\mathbf{X}_n \xrightarrow{d} \lambda'\mathbf{X} \quad \forall \lambda \in \mathbb{R}^k}.$$

A random vector converges in distribution if and only if every linear combination of its elements converges in distribution.

7.3.5 The Continuous Mapping Theorem (Distributional Version)

If $X_n \xrightarrow{d} X$ and $h : \mathbb{R} \to \mathbb{R}$ is continuous (or, more generally, the set of discontinuity points of h has probability zero under the distribution of X), then

$$\boxed{h(X_n) \xrightarrow{d} h(X)}.$$

7.3.6 Slutsky's Theorem

Slutsky's theorem combines convergence in distribution with convergence in probability:

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ (a constant), then

1. $X_n + Y_n \xrightarrow{d} X + c$,
2. $Y_n X_n \xrightarrow{d} cX$,
3. $X_n / Y_n \xrightarrow{d} X/c$ (provided $c \neq 0$).

7.3.7 The Delta Method

The **Delta Method** derives the asymptotic distribution of differentiable functions of asymptotically normal estimators.

Theorem (Univariate Delta Method): If

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2),$$

and $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ with $g'(\theta) \neq 0$, then

$$\boxed{\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2 \sigma^2)}.$$

Multivariate Delta Method: If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(\mathbf{0}, \Sigma)$ and $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable at θ, then

$$\sqrt{n}(\mathbf{g}(\hat{\theta}_n) - \mathbf{g}(\theta)) \xrightarrow{d} N(\mathbf{0}, G\Sigma G'),$$

where $G = \frac{\partial \mathbf{g}}{\partial \theta'}$ is the Jacobian matrix evaluated at θ.

7.3.8 Main Application: The Central Limit Theorem

The most famous application of convergence in distribution is the **Central Limit Theorem**:

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

This justifies the use of the normal distribution for inference (hypothesis tests and confidence intervals) in large samples, even when the original population distribution is unknown.

7.4 Convergence in L_p

7.4.1 Definition of L_p Space and Norm

For $p \geq 1$, the space $L_p(\Omega, \mathcal{F}, \mathbb{P})$ consists of random variables X such that

$$\mathbb{E}[|X|^p] < \infty.$$

The L_p **norm** of a random variable X is

$$\boxed{\|X\|_{L_p} = (\mathbb{E}[|X|^p])^{1/p}}.$$

7.4.2 Definition of L_p Convergence

A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges **in** L_p to X (denoted $X_n \xrightarrow{L_p} X$) if

$$\boxed{\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0}.$$

Case	Name
$p = 1$	Convergence in Mean
$p = 2$	Convergence in Mean Square ($X_n \xrightarrow{m.s.} X$)

7.4.3 Relationships with Other Modes

Relationship	Statement
$L_p \implies$ **Probability**	If $X_n \xrightarrow{L_p} X$, then $X_n \xrightarrow{p} X$ (by Markov's inequality)
Converse	Not true without additional conditions (uniform integrability)
Subsequence Property	If $X_n \xrightarrow{L_p} X$, there exists a subsequence $\{X_{n_j}\}$ such that $X_{n_j} \xrightarrow{a.s.} X$
Order Implication	If $p > q \geq 1$ and $X_n \xrightarrow{L_p} X$, then $X_n \xrightarrow{L_q} X$ (on a probability space)

7.4.4 Important Properties

Property	Description
Completeness (Riesz–Fischer)	L_p spaces are complete metric spaces (Banach spaces)—every Cauchy sequence in L_p converges in L_p
Hölder's Inequality	If $f \in L_p$ and $g \in L_q$ with $1/p + 1/q = 1$, then $fg \in L_1$ and $\mathbb{E}[\lvert fg \rvert] \leq \lVert f \rVert_{L_p} \lVert g \rVert_{L_q}$
Minkowski's Inequality	$\lVert X + Y \rVert_{L_p} \leq \lVert X \rVert_{L_p} + \lVert Y \rVert_{L_p}$
Uniform Integrability	If $X_n \xrightarrow{d} X$ and $\{X_n\}$ is uniformly integrable, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$

7.4.5 Special Case: L_∞

The space L_∞ is defined by the **essential supremum**:

$$\lVert X \rVert_{L_\infty} = \operatorname{ess\,sup} |X| = \inf\{M \geq 0 : \mathbb{P}(|X| > M) = 0\}.$$

Convergence in L_∞ (essentially uniform convergence) implies convergence in L_p for all $p \geq 1$ and convergence in probability. It is the strongest of the L_p modes of convergence.

7.5 Relationships Between Convergence Modes: Summary

7.5.1 Hierarchy of Implications

$$\boxed{\text{Almost Sure} \implies \text{Probability} \implies \text{Distribution}}$$

$$\boxed{L_p\ (p \geq 1) \implies \text{Probability} \implies \text{Distribution}}$$

7.5.2 Special Case: Constant Limit

If the limit is a **constant** c, then

$$\boxed{X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c}.$$

7.5.3 Subsequence Property

If $X_n \xrightarrow{p} X$ (or $X_n \xrightarrow{L_p} X$), then there exists a **subsequence** $\{X_{n_j}\}$ such that

$$X_{n_j} \xrightarrow{a.s.} X.$$

This is a useful result: while convergence in probability does not imply almost sure convergence for the full sequence, it does guarantee the existence of an almost surely convergent subsequence.

7.6 Laws of Large Numbers

7.6.1 Weak Law of Large Numbers (WLLN)

If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. with finite mean $\mu = \mathbb{E}[X]$, then

$$\boxed{\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu}.$$

Versions:

Version	Condition	Proof Method
Bernoulli (1713)	Bernoulli trials	Direct combinatorial argument
Chebyshev (1867)	Finite variance $\mathrm{Var}(X) < \infty$	Chebyshev's inequality
Khinchin (1929)	Finite mean only $\mathbb{E}[\lvert X\rvert] < \infty$	Characteristic functions

The Khinchin version is the most general: it requires only finite absolute mean, no variance assumption.

7.6.2 Strong Law of Large Numbers (SLLN)

If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. with finite mean $\mu = \mathbb{E}[X]$, then

$$\boxed{\bar{X}_n \xrightarrow{a.s.} \mu}.$$

Key Difference: The SLLN guarantees that the **entire trajectory** of the sample mean converges to μ with probability 1. The WLLN only guarantees that the probability of deviation from μ becomes small.

7.6.3 Kolmogorov's Strong Law

The most general form of the SLLN is due to Kolmogorov:

Theorem (Kolmogorov's SLLN): If $\{X_i\}_{i=1}^{\infty}$ are i.i.d., then

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

for some finite constant μ if and only if $\mathbb{E}[\lvert X_1\rvert] < \infty$, in which case $\mu = \mathbb{E}[X_1]$.

7.7 Central Limit Theorems

7.7.1 Classical CLT (Lindeberg–Lévy)

Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. with mean μ and finite variance $\sigma^2 > 0$. Then

$$\boxed{\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)}.$$

Standardized Form:

$$\boxed{Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)}.$$

Implication for Large n:

- $\bar{X}_n \approx N(\mu, \sigma^2/n)$.
- $\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$.

7.7.2 Requirements

Requirement	Description
Independence	Observations must be independent
Identical Distribution	All variables must have the same distribution
Finite Variance	$\sigma^2 < \infty$; this excludes distributions with infinite variance such as the Cauchy distribution

---

7.7.3 Multivariate CLT

Let $\{\mathbf{Y}_i\}_{i=1}^{\infty}$ be i.i.d. random vectors in $\mathbb{R}^k$ with mean $\mu = \mathbb{E}[\mathbf{Y}_i]$ and covariance matrix $\Sigma = \mathrm{Var}(\mathbf{Y}_i)$. Then

$$\boxed{\sqrt{n}(\bar{\mathbf{Y}}_n - \mu) \xrightarrow{d} N(\mathbf{0}, \Sigma)}.$$

Cramér–Wold Device: This can be proved by applying the univariate CLT to every linear combination $\lambda' \mathbf{Y}_i$.

7.7.4 Generalized CLTs

Version	Application
Lindeberg CLT	Independent but **not identically distributed** variables; requires the Lindeberg condition
Lindeberg–Feller CLT	Provides necessary and sufficient conditions for the CLT under independence
Martingale Difference CLT	Time series and econometrics—conditional expectation of error given the past is zero
CLT for Stationary Processes	Serially dependent processes with **long-run covariance matrix** $\Omega = \sum_{h=-\infty}^{\infty} \Gamma(h)$

7.7.5 The Lindeberg Condition

For independent (but not necessarily identically distributed) variables $\{X_{n,i}\}_{i=1}^{k_n}$ with $\mathbb{E}[X_{n,i}] = 0$ and $\sigma_{n,i}^2 = \mathrm{Var}(X_{n,i})$, the **Lindeberg condition** is

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} \frac{1}{s_n^2} \int_{|x|>\epsilon s_n} x^2 \, dF_{n,i}(x) = 0, \quad \forall \epsilon > 0,$$

where $s_n^2 = \sum_{i=1}^{k_n} \sigma_{n,i}^2$.

The Lindeberg condition ensures that no single variable dominates the sum asymptotically. Under this condition, the normalized sum converges in distribution to $N(0,1)$.

7.8 Summary Table of Convergence Modes

Mode	Notation	Definition	Strength
Almost Sure	$X_n \xrightarrow{a.s.} X$	$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$	Strongest
Mean Square (L_2)	$X_n \xrightarrow{m.s.} X$	$\mathbb{E}[(X_n - X)^2] \to 0$	Strong
L_p (General)	$X_n \xrightarrow{L_p} X$	$\mathbb{E}[\lvert X_n - X\rvert^p] \to 0$	Strong
In Probability	$X_n \xrightarrow{p} X$	$\mathbb{P}(\lvert X_n - X\rvert \geq \epsilon) \to 0$	Medium
In Distribution	$X_n \xrightarrow{d} X$	$F_n(x) \to F(x)$ at continuity points	Weakest

7.8.1 Summary of Laws and Theorems

Theorem	Statement	Convergence Mode
WLLN (Chebyshev)	$\bar{X}_n \xrightarrow{p} \mu$ (if $\mathrm{Var}(X) < \infty$)	Probability
WLLN (Khinchin)	$\bar{X}_n \xrightarrow{p} \mu$ (if $\mathbb{E}[\lvert X\rvert] < \infty$)	Probability
SLLN (Kolmogorov)	$\bar{X}_n \xrightarrow{a.s.} \mu$ (if $\mathbb{E}[\lvert X\rvert] < \infty$)	Almost Sure
CLT (Lindeberg–Lévy)	$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$	Distribution
Multivariate CLT	$\sqrt{n}(\bar{\mathbf{X}}_n - \mu) \xrightarrow{d} N(\mathbf{0}, \Sigma)$	Distribution

Chapter 8: Asymptotic Theory

8.1 Order Notation: O and o

8.1.1 Deterministic Order Notation

In asymptotic analysis, the notations O (Big O) and o (small o) characterize the behavior of deterministic sequences in terms of their size or order of magnitude as $n \to \infty$. These definitions are fundamental for describing rates of convergence and the growth of error terms in mathematical models.

8.1.1.1 Big O: Bounded Order of Magnitude The notation **Big O** indicates that a sequence is uniformly bounded by another sequence (or by a constant).

Formal Definition: A sequence of real numbers $\{x_n\}_{n=1}^{\infty}$ is said to be $O(a_n)$ if there exists a constant $M < \infty$ such that

$$\boxed{\frac{|x_n|}{a_n} \leq M \quad \text{for all sufficiently large } n}.$$

Case	Meaning
$x_n = O(1)$	The sequence is **uniformly bounded**: $\exists M$ such that $\lvert x_n\rvert \leq M$ for all sufficiently large n
Interpretation	x_n does not grow faster than the comparison sequence a_n (up to a constant factor)

8.1.1.2 Small o: Negligible Order of Magnitude The notation **small o** describes a sequence that becomes negligible relative to another as $n \to \infty$.

Formal Definition: A sequence $\{x_n\}_{n=1}^{\infty}$ is of order $o(a_n)$ if

$$\boxed{\lim_{n\to\infty} \frac{x_n}{a_n} = 0}.$$

Case	Meaning
$x_n = o(1)$	Equivalent to the sequence **converging to zero**: $x_n \to 0$
Interpretation	x_n grows more slowly than a_n; if a_n is constant, x_n vanishes in the limit

8.1.1.3 Extension to Vectors and Matrices These definitions apply to deterministic vectors and matrices **element-wise**. A matrix A_n is $O(a_n)$ or $o(a_n)$ if each of its individual elements satisfies the respective order condition.

8.1.1.4 Algebra of Order Notation

Rule	Formula
Implication	$x_n = o(a_n) \implies x_n = O(a_n$
Sum	$O(a_n) + O(a_n) = O(a_n)$, $o($
Product	$O(a_n) \cdot O(b_n) = O(a_n b_n)$, o
Composition	$O(o(1)) = o(1)$: a bounded a vanishing sequence vanish
Composition	$o(O(1)) = o(1)$: a sequence constant and vanishes, vani

8.1.1.5 Deterministic Examples

Sequence	Order	Explanation
$x_n = 2 + 1/n$	$O(1)$	Converges to
$x_n = \sqrt{n}$	$O(n^{1/2})$	Grows like t
$x_n = \ln(n)$	$o(n^\ell)$ for any $\ell > 0$	Grows slowe power of n
$x_n = n^2 + 3n$	$O(n^2)$	Dominated b term
$x_n = n^{-1/2}$	$O(n^{-1/2})$ and $o(1)$	Converges to

8.1.2 Stochastic Order Notation (O_p and o_p)

The stochastic definitions extend the deterministic **Big O** and **small o** concepts to random variables. They are fundamental for analyzing the rate of convergence of estimators in probability.

8.1.2.1 Small o in Probability (o_p) **Small o in probability** describes sequences that converge to zero in probability.

Notation	Meaning
$X_n = o_p(1)$	$X_n \xrightarrow{p} 0$ (converges in probability to zero)
$X_n = o_p(a_n)$	$X_n/a_n \xrightarrow{p} 0$ (the ratio converges to zero in probability)
Equivalent	$X_n = a_n \cdot o_p(1)$

8.1.2.2 Big O in Probability (O_p) **Big O in probability** describes sequences that are **stochastically bounded**.

Notation	Meaning
$X_n = O_p(1)$	For every $\epsilon > 0$, there exists $M_\epsilon < \infty$ such that $\mathbb{P}(\lvert X_n \rvert > M_\epsilon) \leq \epsilon$ for all sufficiently large n
Interpretation	The sequence does not "escape" to infinity; it stays within a fixed interval with arbitrarily high probability
$X_n = O_p(a_n)$	$X_n/a_n = O_p(1)$ (the ratio is stochastically bounded)
Equivalent	$X_n = a_n \cdot O_p(1)$

Extension to Vectors and Matrices: A random vector $\mathbf{X}_n$ is $O_p(a_n)$ (or $o_p(a_n)$) if **each individual component** satisfies the respective probability order.

8.1.2.3 Intuition for $O_p(1)$ A sequence $X_n = O_p(1)$ means that for any $\epsilon > 0$, we can find a "fence" M_ϵ such that the probability that X_n falls outside $[-M_\epsilon, M_\epsilon]$ is less than ϵ. The fence may need to be very wide, but it is finite for each ϵ.

Examples: - A sequence that converges in distribution to a random variable is $O_p(1)$. - A sequence that converges in probability to a constant is $O_p(1)$. - A sequence of i.i.d. variables from a distribution with finite mean is $O_p(1)$ (by the WLLN, the average is $O_p(1)$).

8.1.2.4 Relationships between Stochastic Orders

Condition	Implication
Converges in probability to any real number	$\implies O_p(1)$
Converges in distribution to a random variable	$\implies O_p(1)$
Consistent estimator $\hat{\theta}_n$ of θ	$\hat{\theta}_n = \theta + o_p$

8.1.2.5 Asymptotic Algebra Rules The following rules are valid for manipulating stochastic error terms:

Operation	Rule
Sum	$o_p(1) + o_p(1) = o_p(1)$
	$o_p(1) + O_p(1) = O_p(1)$
	$O_p(1) + O_p(1) = O_p(1)$
Product	$o_p(1) \cdot o_p(1) = o_p(1)$
	$o_p(1) \cdot O_p(1) = o_p(1)$ (vanishing term dominates)
	$O_p(1) \cdot O_p(1) = O_p(1)$
Rate Swap	If $X_n = O_p(a_n)$, then $X_n = o_p(b_n)$ for any b_n with $a_n/b_n \to 0$
Example	If $X_n = O_p(n^{-1/2})$, then $X_n = o_p(1)$

8.1.2.6 Relationship with Moments (Hansen's Theorem) The existence of bounded moments guarantees specific probability orders:

Condition	Implication
$\mathbb{E}\lvert X_n \rvert^\delta = O(a_n)$ for some $\delta > 0$	$X_n = O_p(a_n^{1/\delta})$
$\mathbb{E}\lvert X_n \rvert^\delta = o(a_n)$	$X_n = o_p(a_n^{1/\delta})$

In particular: - If $\mathbb{E}[X_n^2] = O(1)$, then $X_n = O_p(1)$. - If $\mathbb{E}[X_n^2] \to 0$, then $X_n = o_p(1)$ (by Chebyshev).

8.2 Slutsky's Theorem

8.2.1 Motivation

Slutsky's Theorem is one of the most frequently used results in asymptotic theory. It allows us to combine sequences that converge in distribution with sequences that converge in probability to constants, greatly simplifying the derivation of the asymptotic distribution of estimators.

Geometric Interpretation: If one sequence is "settling down" to a fixed constant, and another is "settling down" to a distribution, their sum/product/quotient behaves like the sum/product/quotient of the constant and the limiting distribution.

8.2.2 Formal Statement

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables such that $X_n \xrightarrow{d} X$, and let $\{Y_n\}_{n=1}^{\infty}$ be a sequence such that $Y_n \xrightarrow{p} c$, where c is a constant. Then:

1. **Sum:**
$$\boxed{X_n + Y_n \xrightarrow{d} X + c}.$$
2. **Product:**
$$\boxed{Y_n X_n \xrightarrow{d} cX}.$$
3. **Quotient (if $c \neq 0$):**
$$\boxed{\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}}.$$
4. **Inverse (for matrices):** If $A_n \xrightarrow{p} A$ with A non-singular, then
$$\boxed{A_n^{-1} \xrightarrow{p} A^{-1}}.$$

8.2.3 Generalized Slutsky

Let $Y_n \xrightarrow{p} c$ and $X_n \xrightarrow{d} X$. For any function $g(\cdot)$ that is continuous at (c, x) for all x in the support of X,

$$\boxed{g(Y_n, X_n) \xrightarrow{d} g(c, X)}.$$

8.2.4 Applications in Econometrics

Application	Description
Consistent Variance Estimation	If $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$, then $\hat{\sigma}^{-1}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, Q^{-1})$
Asymptotic Standard Errors	Slutsky allows replacing unknown population parameters with consistent estimators without affecting the asymptotic distribution
Wald Tests	The asymptotic distribution of Wald statistics relies on Slutsky to replace the true variance with a consistent estimator

---

8.3 Continuous Mapping Theorem

8.3.1 Motivation

The **Continuous Mapping Theorem (CMT)** states that if a sequence of random variables converges, then any continuous function of that sequence also converges in the same mode. This is essential for deriving the asymptotic distribution of transformations of estimators.

8.3.2 Formal Statement

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors in $\mathbb{R}^d$, and let X be a random vector such that $X_n \xrightarrow{d} X$. Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be continuous on a set C with $\mathbb{P}(X \in C) = 1$. Then

$$\boxed{g(X_n) \xrightarrow{d} g(X)}.$$

Same result holds for convergence in probability and almost sure convergence: - If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$. - If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.

8.3.3 Useful Consequences

Property	Statement
Convergence to Constant	If $X_n \xrightarrow{p} c$ and g is continuous at c, then $g(X_n) \xrightarrow{p} g(c)$
Continuity Almost Everywhere	The CMT only requires continuity on the support of the limiting distribution
Examples	If $\hat{\beta}_n \xrightarrow{p} \beta$, then $1/\hat{\beta}_n \xrightarrow{p} 1/\beta$ (if $\beta \neq 0$), $\ln(\hat{\beta}_n) \xrightarrow{p} \ln(\beta)$ (if $\beta > 0$), etc.

8.3.4 Extension to Stochastic Processes (Donsker's Theorem)

In time series analysis and unit root theory, the CMT is extended to **continuous functionals** of stochastic processes. Donsker's theorem establishes that if a sequence of random functions converges in distribution to a limit process (e.g., Brownian motion), then continuous functionals of these functions—such as integrals, suprema, and supremum norms—also converge in distribution to the corresponding functionals of the limit process.

This is the foundation for the asymptotic distributions of test statistics in unit root and cointegration analysis.

8.4 The Delta Method

8.4.1 Motivation

The **Delta Method** derives the asymptotic distribution of a differentiable function of an asymptotically normal estimator. It is widely used in econometrics to obtain standard errors for non-linear functions of parameters, such as elasticities, marginal effects, odds ratios, and predicted probabilities.

Geometric Interpretation: If we zoom in close enough to the true parameter value, any smooth function looks approximately linear. The Delta Method exploits this linear approximation to transform the asymptotic variance.

8.4.2 Univariate Case

Suppose $\hat{\theta}_n$ is a sequence of estimators such that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2).$$

Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable at θ_0 with $g'(\theta_0) \neq 0$. Then

$$\boxed{\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} N\left(0, \sigma^2[g'(\theta_0)]^2\right)}.$$

Proof Sketch:

1. By Taylor expansion around θ_0,

$$g(\hat{\theta}_n) = g(\theta_0) + g'(\theta_0)(\hat{\theta}_n - \theta_0) + R_n,$$

   where $R_n = o_p(|\hat{\theta}_n - \theta_0|)$.

2. Since $\hat{\theta}_n - \theta_0 = O_p(n^{-1/2})$, we have $R_n = o_p(n^{-1/2})$.

3. Multiplying by $\sqrt{n}$,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) = g'(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1).$$

4. By Slutsky's theorem, the result follows. □

8.4.3 Multivariate Case

Suppose $\hat{\theta}_n$ is a sequence of estimators in $\mathbb{R}^k$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, V).$$

Let $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$ be differentiable at θ_0 with Jacobian matrix

$$G(\theta_0) = \frac{\partial \mathbf{g}}{\partial \theta'}(\theta_0) \in \mathbb{R}^{m \times k}.$$

Then

$$\boxed{\sqrt{n}(\mathbf{g}(\hat{\theta}_n) - \mathbf{g}(\theta_0)) \xrightarrow{d} N\left(\mathbf{0}, G(\theta_0)VG(\theta_0)'\right)}.$$

8.4.4 Special Case: $g(\theta) = \theta^2$

If $g(\theta) = \theta^2$, then $g'(\theta) = 2\theta$. The Delta Method gives:

$$\sqrt{n}(\hat{\theta}_n^2 - \theta_0^2) \xrightarrow{d} N(0, 4\sigma^2\theta_0^2).$$

8.4.5 Applications in Econometrics

Application	Function $g(\theta)$	Use
Elasticity (Log-Log)	$g(\beta) = \beta$ (already linear)	Standard error for elasticity
Semi-Elasticity	$g(\beta) = \beta \cdot \bar{x}$	Standard error for semi-elasticity
Ratio of Coefficients	$g(\beta_1, \beta_2) = \beta_1/\beta_2$	Standard error for odds ratios or cost-benefit ratios
Marginal Effect (Logit)	$g(\beta) = \beta \cdot \Lambda(x'\beta)(1 - \Lambda(x'\beta))$	Standard error for marginal effects in discrete choice models
Predicted Value	$g(\beta) = x_0'\beta$	Standard error for out-of-sample predictions
Exponentiated Coefficient	$g(\beta) = e^{\beta}$	Standard error for odds ratios in logistic regression

---

8.4.6 Important Considerations

Issue	Description
Jacobian Zero	If $G(\theta_0) = \mathbf{0}$, the Delta Method fails; the asymptotic distribution may be non-normal (the "non-standard" case)
Small Sample	For highly non-linear functions, the approximation may be poor in small samples; bootstrap or higher-order approximations may be preferred
Confidence Intervals	The Delta Method is not "transformation invariant"; the confidence interval for $\sqrt{\sigma^2}$ may not equal the square root of the confidence interval for σ^2
Joint Testing	For testing multiple non-linear restrictions, the Wald test uses the multivariate Delta Method

---

8.4.7 The Non-Standard Case: $g'(\theta_0) = 0$

If $g'(\theta_0) = 0$ and $g''(\theta_0) \neq 0$, the asymptotic distribution is quadratic:

$$n(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} \frac{\sigma^2}{2} g''(\theta_0)\chi_1^2.$$

This arises in hypothesis testing when the null imposes a zero gradient.

8.5 Lévy's Continuity Theorem

8.5.1 Motivation

Lévy's Continuity Theorem establishes the equivalence between convergence in distribution and pointwise convergence of characteristic functions. It is a powerful criterion for checking convergence in distribution and is the key tool used in proving the Central Limit Theorem.

8.5.2 Formal Statement

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors in $\mathbb{R}^d$. Let $\phi_n(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}'X_n}]$ be the characteristic function of X_n. Then

$$\boxed{X_n \xrightarrow{d} X \iff \phi_n(\mathbf{t}) \to \phi(\mathbf{t}) \quad \forall \mathbf{t} \in \mathbb{R}^d},$$

where $\phi(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}'X}]$ is the characteristic function of X.

Important: The pointwise limit $\phi(\mathbf{t})$ must be continuous at $\mathbf{t} = \mathbf{0}$ for the implication to hold.

8.5.3 Key Insights

Feature	Description
Sufficiency	If characteristic functions converge pointwise to a function that is continuous at $\mathbf{t} = \mathbf{0}$, then there exists a random vector X with that characteristic function, and $X_n \xrightarrow{d} X$
Necessity	If $X_n \xrightarrow{d} X$, then $\phi_n(\mathbf{t}) \to \phi(\mathbf{t})$ for all $\mathbf{t}$
Advantage	The characteristic function always exists (unlike the MGF, which may not exist for heavy-tailed distributions)
Application	Used to prove the CLT: the characteristic function of the standardized sum converges to $e^{-\|\mathbf{t}\|^2/2}$, the characteristic function of the multivariate normal

---

8.5.4 Moment Generating Function Version

If the moment generating functions exist and are finite, an alternative version states that $X_n \xrightarrow{d} X$ if

$$M_{X_n}(t) \to M_X(t)$$

for all real t in a neighborhood of zero. However, this requires the existence of the MGF, which is not guaranteed for all distributions.

8.5.5 Cramér–Wold Device (Revisited)

Lévy's Continuity Theorem, combined with the Cramér–Wold Device, provides a complete characterization of convergence in distribution for random vectors:

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \lambda'\mathbf{X}_n \xrightarrow{d} \lambda'\mathbf{X} \quad \forall \lambda \in \mathbb{R}^d.$$

This follows from the fact that the characteristic function of $\lambda'\mathbf{X}_n$ is $\phi_n(\lambda)$.

8.6 Summary of Key Results

Concept	Formula
Big O (Deterministic)	$\lvert x_n\rvert/a_n \le M$ for all sufficiently large n
Small o (Deterministic)	$\lim_{n\to\infty} x_n/a_n = 0$
Big O in Probability	$\forall\epsilon > 0, \exists M_\epsilon : \mathbb{P}(\lvert X_n\rvert > M_\epsilon) \le \epsilon$
Small o in Probability	$X_n/a_n \xrightarrow{p} 0$
Slutsky's Theorem	$X_n \xrightarrow{d} X,\ Y_n \xrightarrow{p} c \implies X_n + Y_n \xrightarrow{d} X + c,\ Y_n X_n \xrightarrow{d} cX$
Continuous Mapping	$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$ (for continuous g)
Delta Method (Univariate)	$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} N(0, \sigma^2[g'(\theta_0)]^2)$
Delta Method (Multivariate)	$\sqrt{n}(\mathbf{g}(\hat{\theta}_n) - \mathbf{g}(\theta_0)) \xrightarrow{d} N(\mathbf{0}, GVG')$
Lévy's Continuity	$X_n \xrightarrow{d} X \iff \phi_n(\mathbf{t}) \to \phi(\mathbf{t})\ \forall\mathbf{t}$
Cramér–Wold	$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \lambda'\mathbf{X}_n \xrightarrow{d} \lambda'\mathbf{X}\ \forall\lambda$

Chapter 9: Point Estimation

9.1 Estimator and Estimate

9.1.1 Motivation

The central problem that point estimation solves is the **mapping of unknown population characteristics** (parameters) from a finite sample. Since it is infeasible or destructive to observe the entire population in many contexts—such as durability testing or quality control—we need a mathematical rule that processes sample data and returns an educated "guess" about the true parameter value.

Geometric Interpretation in $\mathbb{R}^n$: Visualize the sample as a vector $\mathbf{y} \in \mathbb{R}^n$ in an n-dimensional observation space. In least squares estimation, the goal is to find a parameter vector β such that the linear combination $\mathbf{X}\beta$ is as close as possible to $\mathbf{y}$. Geometrically, the fitted value $\hat{\mathbf{y}}$ is the **orthogonal projection** of $\mathbf{y}$ onto the column space of $\mathbf{X}$. The estimator is the projection rule; the estimate is the exact coordinate of the projected point for a specific dataset.

9.1.2 Formal Definitions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{X_i\}_{i=1}^n$ an independent and identically distributed (i.i.d.) sample with probability density function $f(x;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is an unknown parameter belonging to the parameter space Θ.

1. **Estimator ($\hat{\theta}_n$):** A statistic $h : \mathbb{R}^n \to \Theta$, defined as a measurable function of the random sample:

$$\boxed{\hat{\theta}_n = h(X_1, X_2, \ldots, X_n)}.$$

   The estimator is a **random variable** with its own sampling distribution.

2. **Estimate ($\hat{\theta}_{obs}$):** The numerical value realized by the estimator when applied to a specific realization of the sample $(x_1, x_2, \ldots, x_n)$:

$$\boxed{\hat{\theta}_{obs} = h(x_1, x_2, \ldots, x_n)}.$$

   The estimate is a **numerical constant** (a point in $\mathbb{R}^k$).

9.1.3 Mean Squared Error Decomposition

Theorem (Decomposition of Mean Squared Error): The Mean Squared Error (MSE) of an estimator $\hat{\theta}_n$ for a parameter θ can be decomposed as

$$\boxed{\text{MSE}(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n) + [\text{Bias}(\hat{\theta}_n)]^2}.$$

Proof:

1. By definition, $\text{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$.
2. Add and subtract $\mathbb{E}[\hat{\theta}_n]$:
$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta)^2].$$
3. Expanding the square:
$$\text{MSE} = \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + 2(\mathbb{E}[\hat{\theta}_n] - \theta)\mathbb{E}[\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]] + (\mathbb{E}[\hat{\theta}_n] - \theta)^2.$$
4. The cross term vanishes since $\mathbb{E}[\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]] = 0$. Therefore,
$$\text{MSE}(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n) + [\text{Bias}(\hat{\theta}_n)]^2.$$

□

9.1.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Finite First Moment $\mathbb{E}[\lvert X\rvert] < \infty$	Ensures $\mathbb{E}[\hat{\theta}_n]$ is well-defined	Cauchy distribution; mean undefined
Finite Second Moment $\mathbb{E}[X^2] < \infty$	Ensures $\text{Var}(\hat{\theta}_n)$ and MSE are finite	Pareto with $\alpha \leq 2$; infinite variance
Parameter Identifiability	$\theta_1 \neq \theta_2 \implies f(x;\theta_1) \neq f(x;\theta_2)$	Perfect multicollinearity in regression

---

9.2 Bias and Mean Squared Error

9.2.1 Motivation

The concept of bias addresses the problem of **systematic miscalibration**. An estimator can have extremely low variance (be very precise) but consistently "aim" at the wrong target. Bias quantifies this average discrepancy between the expected value of the estimator and the true parameter.

Geometric Interpretation: In $\mathbb{R}$, bias is the distance between the center of mass of the estimator's sampling distribution and the true parameter θ. The MSE resolves the **bias-variance trade-off**: it is the squared L^2 distance between the estimator and the target.

9.2.2 Formal Definitions

Let $\hat{\theta}_n = h(X_1, \ldots, X_n)$ be an estimator for $\theta \in \Theta$.

1. **Bias:**
$$\boxed{\text{Bias}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[\hat{\theta}_n] - \theta}.$$
   - $\hat{\theta}_n$ is **unbiased** if $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$ for all $\theta \in \Theta$.
2. **Mean Squared Error:**
$$\boxed{\text{MSE}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2]}.$$

9.2.3 Bias of the Sample Variance Estimator

Theorem: Let $\{X_i\}_{i=1}^n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. The estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is biased for σ^2, with

$$\boxed{\text{Bias}(\hat{\sigma}^2) = -\frac{\sigma^2}{n}}.$$

Proof:

1. Using the identity $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$.
2. Taking expectations:

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2.$$

3. Therefore,

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2.$$

4. Hence,

$$\text{Bias}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

□

Remark: The unbiased estimator is $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

9.2.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Finite First Moment $\mathbb{E}[\lvert X\rvert] < \infty$	Bias $\mathbb{E}[\hat{\theta}] - \theta$ well-defined	Cauchy distribution
Finite Second Moment $\mathbb{E}[X^2] < \infty$	MSE finite	Pareto with $\alpha \leq 2$

9.3 Method of Moments

9.3.1 Motivation

The Method of Moments (MM), formalized by Karl Pearson in 1894, addresses the problem of **computational complexity** by transforming estimation into a system of algebraic equations based on sample moments. The principle is simple: if population moments are functions of unknown parameters, equate them to their sample counterparts.

Geometric Interpretation: Visualize the MM as a mapping $\Phi : \Theta \to \mathcal{M}$ from the parameter space to the moment space. In the just-identified case, the estimator $\hat{\theta}$ is the point whose image $\Phi(\hat{\theta})$ exactly coincides with the vector of sample moments $\mathbf{m}_n$.

9.3.2 Formal Definitions

Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from $f(x;\theta)$, where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

1. **Population Moments:**
$$\mu_r'(\theta) = \mathbb{E}[X^r] = \int_{-\infty}^{\infty} x^r f(x;\theta)\, dx, \quad r = 1, 2, \ldots$$

2. **Sample Moments:**
$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, \quad r = 1, 2, \ldots$$

3. **Principle of Analogy:** The MM estimator $\hat{\theta}_n$ solves the system of k equations:
$$\boxed{\mu_r'(\hat{\theta}_n) = m_r, \quad r = 1, 2, \ldots, k}.$$

9.3.3 Consistency of the MM Estimator

Theorem (Consistency of MM): Under regularity conditions (finite moments and continuous invertible mapping), the MM estimator is consistent:

$$\boxed{\hat{\theta}_n \xrightarrow{p} \theta_0}.$$

Proof Sketch:

1. Let $\mathbf{G}(\theta) = (\mu_1'(\theta), \ldots, \mu_k'(\theta))'$.
2. By construction, $\mathbf{G}(\hat{\theta}_n) = \mathbf{m}_n$.
3. By the WLLN, $\mathbf{m}_n \xrightarrow{p} \mathbf{G}(\theta_0)$.
4. If $\mathbf{G}$ has a continuous inverse, then
$$\hat{\theta}_n = \mathbf{G}^{-1}(\mathbf{m}_n) \xrightarrow{p} \mathbf{G}^{-1}(\mathbf{G}(\theta_0)) = \theta_0.$$

□

9.3.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Finite Moments $\mathbb{E}[\lvert X\rvert^k] < \infty$	Population moments well-defined	Pareto with $\alpha \leq k$
Invertibility of Moment Function	Unique solution to moment equations	Symmetric parameterization $(\theta_1 + \theta_2)$

9.4 Method of Moments — Applications

9.4.1 MM Estimators for the Gamma Distribution

Let $\{X_i\}_{i=1}^n$ be i.i.d. from Gamma(α, β) with density

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x \geq 0,$$

where $\alpha > 0$ (shape) and $\beta > 0$ (scale).

Theoretical Moments: - $\mu_1' = \mathbb{E}[X] = \alpha\beta$ - $\mu_2' = \mathbb{E}[X^2] = \alpha\beta^2 + (\alpha\beta)^2 = \alpha(\alpha+1)\beta^2$

Sample Moments: - $m_1 = \bar{X}$ - $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$

9.4.2 MM Estimators for Gamma

Theorem: The method of moments estimators for the Gamma distribution are

$$\boxed{\hat{\alpha}_{MM} = \frac{\bar{X}^2}{\hat{\sigma}^2}}, \qquad \boxed{\hat{\beta}_{MM} = \frac{\hat{\sigma}^2}{\bar{X}}},$$

where $\hat{\sigma}^2 = m_2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Proof:

1. By the principle of analogy:
$$\alpha\beta = \bar{X}, \quad \alpha(\alpha+1)\beta^2 = m_2.$$
2. From $\alpha\beta = \bar{X}$, we have $\alpha = \bar{X}/\beta$.
3. Substitute into the second equation:
$$\frac{\bar{X}}{\beta}\left(\frac{\bar{X}}{\beta} + 1\right)\beta^2 = m_2 \implies \bar{X}^2 + \bar{X}\beta = m_2.$$
4. Since $m_2 = \bar{X}^2 + \hat{\sigma}^2$, we get $\bar{X}\beta = \hat{\sigma}^2$.
5. Therefore, $\hat{\beta}_{MM} = \hat{\sigma}^2/\bar{X}$ and $\hat{\alpha}_{MM} = \bar{X}^2/\hat{\sigma}^2$. □

9.4.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Finite Moments $\mathbb{E}[X^2] < \infty$	Sample variance well-defined	Pareto with $\alpha \leq 2$
i.i.d. Sampling	Sample moments converge to population moments	Autocorrelated time series

---

9.5 Asymptotic Properties of the Method of Moments

9.5.1 Motivation

Asymptotic properties answer: what happens to our estimator as $n \to \infty$? **Consistency** ensures the estimation error vanishes. **Asymptotic normality** enables hypothesis testing and confidence interval construction, even when the finite-sample distribution is unknown.

Geometric Interpretation: Visualize Θ as the parameter space. **Consistency** is the collapse of the estimator's probability mass onto θ_0. **Asymptotic normality** describes the "cloud" of uncertainty: as the cloud shrinks, its shape becomes a symmetric multidimensional ellipse.

9.5.2 Formal Statement

Let $\hat{\theta}_n$ be the MM estimator satisfying $\mathbf{m}_n = \mu(\hat{\theta}_n)$, where: - $\mathbf{m}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(X_i)$ (sample moments) - $\mu(\theta) = \mathbb{E}[\mathbf{g}(X)]$ (population moments)

Theorem: Under regularity conditions (finite second moments, μ continuously differentiable with full-rank Jacobian),

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})},$$

where

$$\mathbf{V} = [\mathbf{G}'\mathbf{G}]^{-1}\mathbf{G}'\Sigma\mathbf{G}[\mathbf{G}'\mathbf{G}]^{-1},$$

with $\mathbf{G} = \left.\frac{\partial\mu(\theta)}{\partial\theta'}\right|_{\theta_0}$ and $\Sigma = \mathrm{Var}(\mathbf{g}(X))$.

9.5.3 Proof Sketch

1. By Taylor expansion around θ_0:
$$\mathbf{G}(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) = \mathbf{m}_n - \mu(\theta_0).$$
2. Multiply by $\sqrt{n}$:
$$\mathbf{G}(\bar{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}(\mathbf{m}_n - \mu(\theta_0)).$$
3. By the CLT, $\sqrt{n}(\mathbf{m}_n - \mu(\theta_0)) \xrightarrow{d} N(\mathbf{0}, \Sigma)$.
4. By the WLLN, $\mathbf{G}(\bar{\theta}_n) \xrightarrow{p} \mathbf{G}(\theta_0) = \mathbf{G}$.
5. By Slutsky's theorem, the result follows. □

9.5.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
WLLN	Sample moments converge to population moments	Non-stationary random walk
Differentiability	Taylor expansion valid	Non-smooth moment function
Full Rank Jacobian	$\mathbf{G}$ invertible	Parameter redundancy

---

9.6 The Likelihood Function

9.6.1 Motivation

The likelihood function solves the problem of **logical inversion of the probabilistic process**. While probability predicts data $\mathbf{x}$ given θ, likelihood uses observed data to assess how "compatible" different candidate values of θ are.

Geometric Interpretation: The likelihood function $L_n(\theta)$ defines a **surface** over the parameter space Θ. Unlike a pdf, which integrates to 1 over the sample space, the likelihood has no such constraint in the parameter space. MLE consists of finding the **global maximum** of this surface.

9.6.2 Formal Definitions

Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample with density $f(x_i;\theta)$, $\theta \in \Theta$.

1. **Likelihood Function:**
$$\boxed{L_n(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta)}.$$
2. **Log-Likelihood Function:**
$$\boxed{\ell_n(\theta) = \ln L_n(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)}.$$

9.6.3 Equivalence of Maximization

Theorem: The value of θ that maximizes $L_n(\theta)$ is identical to the value that maximizes $\ell_n(\theta)$.

Proof: Since $\ln(\cdot)$ is strictly increasing, the maximizer is preserved under the logarithmic transformation. □

9.6.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
i.i.d. Sampling	Joint density factors into product of marginals	Autocorrelated time series
Support Independent of θ	Likelihood differentiable everywhere	Uniform $(0, \theta)$ distribution

---

9.7 The Score Function and Hessian

9.7.1 Formal Definitions

Let $\ell_n(\theta) = \sum_{i=1}^{n} \ln f(X_i; \theta)$ be the log-likelihood.

1. **Score Function:** The gradient of the log-likelihood:

$$\boxed{S_n(\theta) = \nabla_\theta \ell_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i; \theta)}.$$

2. **Hessian:** The matrix of second partial derivatives:

$$\boxed{H_n(\theta) = \nabla_\theta \nabla_\theta' \ell_n(\theta)}.$$

9.7.2 Zero Expectation of the Score

Theorem: Under regularity conditions,

$$\boxed{\mathbb{E}_{\theta_0}[S_n(\theta_0)] = \mathbf{0}}.$$

Proof:

1. By definition,
$$\mathbb{E}[S_n(\theta_0)] = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} \ln f(x; \theta_0) f(x; \theta_0)\, dx.$$

2. Since $\frac{\partial}{\partial \theta} \ln f = \frac{1}{f} \frac{\partial f}{\partial \theta}$,
$$\mathbb{E}[S_n(\theta_0)] = \int_{\mathcal{X}} \frac{\partial f(x; \theta_0)}{\partial \theta}\, dx.$$

3. By the regularity condition (interchange of integral and derivative),
$$\mathbb{E}[S_n(\theta_0)] = \frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x; \theta_0)\, dx = \frac{\partial}{\partial \theta}(1) = \mathbf{0}.$$

□

9.8 Maximum Likelihood Estimation — Bernoulli and Normal

9.8.1 MLE for Bernoulli

Let $\{X_i\}_{i=1}^{n}$ be i.i.d. with $X_i \sim \text{Bernoulli}(p)$, $p \in (0, 1)$.

Theorem: The MLE of p is

$$\boxed{\hat{p}_{MLE} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i}.$$

Proof:

1. The likelihood is $L_n(p) = p^{\sum X_i}(1-p)^{n-\sum X_i}$.
2. The log-likelihood is $\ell_n(p) = (\sum X_i)\ln p + (n - \sum X_i)\ln(1-p)$.
3. The score is $S_n(p) = \frac{\sum X_i}{p} - \frac{n-\sum X_i}{1-p}$.
4. Setting $S_n(p) = 0$: $\frac{\sum X_i}{\hat{p}} = \frac{n-\sum X_i}{1-\hat{p}}$.
5. Solving: $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$. The second derivative is strictly negative, confirming a maximum. □

9.8.2 MLE for Normal

Let $\{X_i\}_{i=1}^n$ be i.i.d. with $X_i \sim N(\mu, \sigma^2)$.

Theorem: The MLEs for the Normal parameters are

$$\boxed{\hat{\mu}_{MLE} = \bar{X}}, \qquad \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Proof:

1. The log-likelihood is

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

2. Differentiate with respect to μ:

$$\frac{\partial \ell_n}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = 0 \implies \hat{\mu} = \bar{X}.$$

3. Differentiate with respect to σ^2:

$$\frac{\partial \ell_n}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (X_i - \mu)^2 = 0.$$

4. Substituting $\hat{\mu} = \bar{X}$:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

□

9.8.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
i.i.d. Sampling	Likelihood factors into product	Correlated observations
Parameter in Interior	Maximum found by setting gradient to zero	True $p = 0$ or $p = 1$

9.9 Asymptotic Properties of the MLE

9.9.1 Formal Statement

Let $\hat{\theta}_n$ be the MLE of $\theta_0 \in \Theta \subseteq \mathbb{R}^k$. Under regularity conditions:

1. **Consistency:**
$$\boxed{\hat{\theta}_n \xrightarrow{p} \theta_0}.$$
2. **Asymptotic Normality:**
$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\theta_0)^{-1})},$$
where $\mathcal{I}(\theta_0)$ is the Fisher Information matrix per observation.
3. **Asymptotic Efficiency:** The MLE achieves the Cramér–Rao lower bound asymptotically.

9.9.2 Proof Sketch of Asymptotic Normality

1. The MLE satisfies the first-order condition: $S_n(\hat{\theta}_n) = \mathbf{0}$.
2. By Taylor expansion around θ_0:
$$S_n(\hat{\theta}_n) \approx S_n(\theta_0) + H_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) = \mathbf{0}.$$
3. Rearranging:
$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left[-\frac{1}{n} H_n(\bar{\theta}_n)\right]^{-1} \left[\frac{1}{\sqrt{n}} S_n(\theta_0)\right].$$
4. By the CLT, $\frac{1}{\sqrt{n}} S_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\theta_0))$.
5. By the WLLN, $-\frac{1}{n} H_n(\bar{\theta}_n) \xrightarrow{p} \mathcal{I}(\theta_0)$.
6. By Slutsky's theorem, the result follows. □

9.9.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Global Identifiability	Unique maximizer of expected likelihood	Perfect multicollinearity
Support Independence	Interchange of integration and differentiation	Uniform $[0, \theta]$ distribution

---

9.10 Fisher Information Matrix

9.10.1 Formal Definitions

Let $X \sim f(x; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Define: - $s(\theta; x) = \nabla_\theta \ln f(x; \theta)$ (Score vector) - $H(\theta; x) = \nabla_\theta \nabla_\theta' \ln f(x; \theta)$ (Hessian)

Fisher Information Matrix:
$$\boxed{\mathcal{I}(\theta) = \mathbb{E}_\theta[s(\theta; X) s(\theta; X)']}.$$

9.10.2 Information Matrix Equality

Theorem: Under regularity conditions,

$$\boxed{\mathcal{I}(\theta) = -\mathbb{E}_\theta[H(\theta; X)]}.$$

Proof:

1. From $\int f(x;\theta)\,dx = 1$, differentiate twice:

$$\int [H(\theta; x)f(x;\theta) + s(\theta; x)s(\theta; x)' f(x;\theta)]\,dx = \mathbf{0}.$$

2. Therefore,

$$\mathbb{E}[H(\theta; X)] + \mathbb{E}[s(\theta; X)s(\theta; X)'] = \mathbf{0}.$$

3. Hence, $\mathcal{I}(\theta) = -\mathbb{E}[H(\theta; X)]$. □

9.10.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Second-Order Differentiability	Hessian exists	Laplace distribution
Support Independence	Interchange of integration and differentiation	Uniform $[0, \theta]$

9.11 Cramér–Rao Inequality

9.11.1 Formal Statement

Theorem (Cramér–Rao Lower Bound): Let $W(\mathbf{X})$ be an unbiased estimator of θ. Under regularity conditions,

$$\boxed{\mathrm{Var}_\theta(W(\mathbf{X})) \geq \frac{1}{n\mathcal{I}(\theta)}}.$$

9.11.2 Proof

1. For an unbiased estimator, $\int W(\mathbf{x})L(\theta; \mathbf{x})\,d\mathbf{x} = \theta$.
2. Differentiate with respect to θ:

$$\int W(\mathbf{x})S_n(\theta)L(\theta; \mathbf{x})\,d\mathbf{x} = 1.$$

3. Thus, $\mathbb{E}_\theta[WS_n] = 1$.
4. By Cauchy–Schwarz,

$$[\mathrm{Cov}(W, S_n)]^2 \leq \mathrm{Var}(W)\mathrm{Var}(S_n).$$

5. Since $\mathrm{Cov}(W, S_n) = 1$ and $\mathrm{Var}(S_n) = n\mathcal{I}(\theta)$,

$$1 \leq \mathrm{Var}(W) \cdot n\mathcal{I}(\theta).$$

6. Therefore, $\mathrm{Var}(W) \geq \frac{1}{n\mathcal{I}(\theta)}$. □

9.11.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Unbiasedness	CRLB applies only to unbiased estimators	James–Stein estimator (biased, lower variance)
Support Independence	Differentiation under integral sign	Uniform $[0, \theta]$

9.12 Invariance Property of the MLE

9.12.1 Formal Statement

Theorem (Invariance of the MLE): If $\hat{\theta}_n$ is the MLE of θ and $\eta = h(\theta)$ is a continuous transformation, then $\hat{\eta}_n = h(\hat{\theta}_n)$ is the MLE of η.

9.12.2 Proof

1. Define the induced likelihood: $L^*(\eta) = \sup_{\{\theta : h(\theta) = \eta\}} L(\theta)$.
2. Since $\hat{\theta}$ is the global maximizer, $L(\hat{\theta}) \geq L(\theta)$ for all θ.
3. Therefore, $L^*(h(\hat{\theta})) = L(\hat{\theta})$.
4. For any other η, $L^*(\eta) \leq L(\hat{\theta}) = L^*(h(\hat{\theta}))$.
5. Hence, $h(\hat{\theta})$ maximizes $L^*(\eta)$. □

9.12.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Existence of MLE	$\hat{\theta} \in \Theta$	Monotonic likelihood with no maximum
Measurability of h	$h(\hat{\theta})$ is a random variable	Non-measurable function

---

9.13 Asymptotic Efficiency of the MLE

9.13.1 Formal Definition

An estimator $\hat{\theta}_n$ is **asymptotically efficient** if it is consistent, asymptotically normal, and its asymptotic variance equals the Cramér–Rao lower bound:

$$\boxed{\mathbf{V} = \mathcal{I}(\theta_0)^{-1}}.$$

9.13.2 Theorem

Theorem: Under regularity conditions, the MLE is asymptotically efficient.

Proof: Follows directly from the asymptotic normality result:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}).$$

□

9.13.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Correct Model Specification	MLE achieves CRLB	Misspecified model (QMLE)
Invertible Information Matrix	Asymptotic variance finite	Parameter redundancy

9.14 Summary of Key Results

Concept	Definition	Key Formula
Estimator	Function of the sample	$\hat{\theta}_n = h(X_1, \ldots, X_n)$
Estimate	Numerical value for a specific sample	$\hat{\theta}_{obs} = h(x_1, \ldots, x_n)$
Bias	$\mathbb{E}[\hat{\theta}_n] - \theta$	$\text{Bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$
MSE	$\mathbb{E}[(\hat{\theta}_n - \theta)^2]$	$\text{MSE} = \text{Var}(\hat{\theta}_n) + [\text{Bias}(\hat{\theta}_n)]^2$
Method of Moments	Equate population and sample moments	$\mu'_r(\hat{\theta}_n) = m_r$
Likelihood	Joint density as function of θ	$L_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$
Score	Gradient of log-likelihood	$S_n(\theta) = \nabla_\theta \ell_n(\theta)$
Fisher Information	Variance of the score	$\mathcal{I}(\theta) = \mathbb{E}[s(\theta)s(\theta)'] = -\mathbb{E}[H(\theta)]$
Cramér–Rao Bound	Lower bound for unbiased estimators	$\text{Var}(\hat{\theta}_n) \geq 1/[n\mathcal{I}(\theta)]$
MLE Asymptotic Normality	$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$	Efficient estimator
Invariance of MLE	MLE of $h(\theta)$ is $h(\hat{\theta}_n)$	$\hat{\eta}_n = h(\hat{\theta}_n)$
Asymptotic Efficiency	Achieves CRLB asymptotically	$\mathbf{V} = \mathcal{I}(\theta_0)^{-1}$

Chapter 10: Hypothesis Testing

10.1 Fundamental Concepts of Hypothesis Testing

10.1.1 Motivation

The fundamental problem that **hypothesis testing** solves is decision-making under statistical uncertainty. Researchers often need to choose between two conflicting theories or states of nature based solely on a finite sample of data. Concrete examples include deciding whether a new vaccine is effective, whether education affects average salary, or whether a coin is fair.

Geometric Interpretation: Hypothesis testing can be interpreted as a partition of two spaces:

1. **Parameter Space (Θ):** Imagine a plane $\mathbb{R}^2$ where the axes represent two parameters, such as the mean μ and variance σ^2. The **null hypothesis (H_0)** defines a specific region Θ_0 (e.g., a line where $\mu = \mu_0$), while the **alternative hypothesis (H_1)** defines its complement $\Theta_1 = \Theta \setminus \Theta_0$.
2. **Sample Space ($\mathcal{X}$):** If we observe n variables, each sample is a point in $\mathbb{R}^n$. Hypothesis testing defines a **critical region** (rejection region) $R \subset \mathcal{X}$ and an acceptance region R^c. If the observed sample point falls within R, we reject H_0.

10.1.2 Formal Definitions

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector on a sample space $\mathcal{X} \subseteq \mathbb{R}^n$, with distribution $F(x; \theta)$ indexed by $\theta \in \Theta \subseteq \mathbb{R}^k$. A **hypothesis test** is a decision rule based on a partition of the parameter space into two disjoint subsets:

$$\Theta_0 \cap \Theta_1 = \emptyset \quad \text{and} \quad \Theta_0 \cup \Theta_1 = \Theta.$$

The **null hypothesis** is

$$\boxed{H_0 : \theta \in \Theta_0}.$$

The **alternative hypothesis** is

$$\boxed{H_1 : \theta \in \Theta_1}.$$

The test is operationalized by a **test function** $\phi : \mathcal{X} \to \{0, 1\}$, where $\phi(x) = 1$ indicates rejection of H_0 and $\phi(x) = 0$ indicates non-rejection. The **rejection region** is

$$R = \{x \in \mathcal{X} : \phi(x) = 1\}.$$

10.1.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Exhaustiveness and Mutual Exclusivity	Ensures a clear decision rule	$H_0 : \mu \leq 10$, $H_1 : \mu \geq 10$ overlap at $\mu = 10$
Identifiability and Correct Specification	Distribution of test statistic determinable	Testing mean under normality when data are Cauchy
Random Sampling (i.i.d.)	Justifies limit theorems for test distribution	Strongly autocorrelated time series; spurious regression

---

10.2 Test Statistic and Critical Region

10.2.1 Motivation

The central objective is **dimensionality reduction** for statistical decision-making. A raw sample is a point in $\mathbb{R}^n$, making decision boundaries complex. The test statistic compresses the sample into a scalar, simplifying the decision to a one-dimensional threshold.

Geometric Interpretation:

1. **Projection Mapping:** The **Test Statistic** acts as $T : \mathbb{R}^n \to \mathbb{R}$, compressing all relevant sample information into a single scalar.
2. **Partition of the Sample Space:** In $\mathbb{R}^n$, the **Critical Region** is $R \subset \mathbb{R}^n$, partitioning the sample space into acceptance and rejection regions.
3. **Visualization in $\mathbb{R}$:** After applying T, the decision becomes one-dimensional: $R = \{t \in \mathbb{R} : t > c\}$, where c is the critical value.

10.2.2 Formal Definitions

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with distribution indexed by θ. Consider $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$.

A **Test Statistic** is a measurable function $T : \mathcal{X} \to \mathbb{R}$. The **Critical Region** is

$$R = \{t \in \mathrm{Im}(T) : t > c\},$$

where c is the **critical value**.

The decision rule is

$$\phi(x) = \mathbf{1}_R(T(x)) = \begin{cases} 1, & \text{if } T(x) \in R, \\ 0, & \text{if } T(x) \notin R, \end{cases}$$

where $\phi(x) = 1$ implies rejection of H_0.

10.2.3 Determination of the Critical Value

Theorem: Let $G_0(t) = \mathbb{P}(T(X) \leq t \mid H_0)$ be the CDF of T under the null. If G_0 is continuous and strictly increasing, the critical value c for a test of size α is

$$\boxed{c = G_0^{-1}(1-\alpha)}.$$

Proof:

1. By definition, $\alpha = \mathbb{P}(\text{Reject } H_0 \mid H_0) = \mathbb{P}(T(X) > c \mid H_0)$.
2. By the complement rule, $\alpha = 1 - G_0(c)$.
3. Thus, $G_0(c) = 1 - \alpha$, so $c = G_0^{-1}(1-\alpha)$. $\square$

10.2.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Measurability of T	$T(X)$ is a well-defined random variable	T maps to non-measurable sets
Pivotal Distribution under H_0	Critical value does not depend on unknown parameters	Distribution of T depends on unknown σ^2
Monotonicity of Likelihood Ratio	Larger T means stronger evidence against H_0	Pathological multimodal distributions

---

10.3 Type I and Type II Errors

10.3.1 Motivation

Since inference is based on finite samples, every decision is subject to uncertainty. We can fail in two logically distinct ways: (i) rejecting a correct theory or (ii) accepting a false theory. The formalization of **Type I and Type II Errors** allows the researcher to quantify these risks and establish a balance between them.

Geometric Interpretation: Imagine the pdfs of the test statistic T under H_0 and a specific alternative H_1. Given a critical value c: - The area under H_0 in the rejection region is α (Type I Error). - The area under H_1 in the acceptance region is β (Type II Error). - Moving c right decreases α but increases β—the fundamental trade-off.

10.3.2 Formal Definitions

Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Let R be the critical region.

1. **Type I Error:**
$$\boxed{\alpha = \mathbb{P}(T(X) \in R \mid \theta = \theta_0)}.$$
2. **Type II Error:**
$$\boxed{\beta = \mathbb{P}(T(X) \notin R \mid \theta = \theta_1)}.$$
3. **Power of the Test:**
$$\boxed{\pi(\theta_1) = 1 - \beta = \mathbb{P}(T(X) \in R \mid \theta = \theta_1)}.$$

10.3.3 The Trade-off between α and β

Theorem: For a fixed sample size n, decreasing α increases β.

Proof (Normal mean, known variance):

1. Let $X_i \sim N(\mu, \sigma^2)$, σ^2 known. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1 > \mu_0$. Reject if $\bar{X} > c$.
2. By definition,
$$\alpha = \mathbb{P}(\bar{X} > c \mid \mu = \mu_0) \implies c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$
3. The Type II error is
$$\beta = \mathbb{P}(\bar{X} \leq c \mid \mu = \mu_1) = \Phi\left(z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right).$$
4. Since Φ is strictly increasing, decreasing α (increasing z_α) increases β. $\square$

10.3.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
i.i.d. Sampling	Variance of test statistic decreases with n	Autocorrelated data inflate Type I error
Fixed Alternative for β	β depends on how false H_0 is	Composite alternative; power is a function
Significance $\alpha > 0$	Test has capacity to reject	$\alpha = 0 \implies \beta = 1$ for all alternatives

---

10.4 Power Function

10.4.1 Motivation

The **Power Function** solves the problem of global test evaluation across the entire parameter space. While α focuses exclusively on the risk under the null, the power function allows visualization of the rejection probability for any parameter value, determining the test's **sensitivity**.

Geometric Interpretation: The power function $\pi(\theta)$ maps $\Theta \to [0, 1]$. For a well-behaved test, the curve rises from α at θ_0 toward 1 as θ moves away from θ_0. The "slope" of this curve defines the test's **accuracy**.

10.4.2 Formal Definition

Let $R \subset \mathcal{X}$ be the critical region for $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. The **Power Function** is

$$\boxed{\pi(\theta) = \mathbb{P}_\theta(X \in R), \quad \forall \theta \in \Theta}.$$

- If $\theta \in \Theta_0$, $\pi(\theta)$ is the Type I error probability; $\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha$ is the size.
- If $\theta \in \Theta_1$, $\pi(\theta) = 1 - \beta(\theta)$ is the power.

10.4.3 Power Function for Normal Means

Theorem: Let $X_i \sim N(\theta, \sigma^2)$ i.i.d. with σ^2 known. For $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$ at level α,

$$\boxed{\pi(\theta) = \Phi\left(\frac{\sqrt{n}(\theta - \theta_0)}{\sigma} - z_{1-\alpha}\right)}.$$

Proof:

1. The rejection region is $R = \{\bar{X} > c\}$, with $c = \theta_0 + z_{1-\alpha}\sigma/\sqrt{n}$.
2. By definition, $\pi(\theta) = \mathbb{P}_\theta(\bar{X} > c)$.
3. Standardizing: $\pi(\theta) = \mathbb{P}\left(Z > z_{1-\alpha} - \frac{\sqrt{n}(\theta-\theta_0)}{\sigma}\right)$.
4. By symmetry: $\pi(\theta) = \Phi\left(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} - z_{1-\alpha}\right)$. □

10.4.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Continuity of CDF	Power function continuous in θ	Discrete distributions; power has jumps
Consistency of Test	Power $\to 1$ as $n \to \infty$ for $\theta \neq \theta_0$	Statistic with non-vanishing variance
Unimodality of Likelihood Ratio	Power increases monotonically as θ moves away from θ_0	Mixture models with non-monotonic power

---

10.5 Likelihood Ratio Test (LRT)

10.5.1 Motivation

The **Likelihood Ratio Test (LRT)** solves the problem of testing hypotheses where the restriction imposed by H_0 defines a subspace of the parameter space. The central question is whether the loss of fit when imposing a restriction is statistically significant or merely due to sampling variability.

Geometric Interpretation: Imagine the log-likelihood $\ell(\theta)$ as a "mountain" over Θ. The unrestricted MLE $\hat{\theta}$ is the peak. The null hypothesis defines a subset Θ_0 (e.g., a plane cutting through the mountain). The restricted MLE $\tilde{\theta}$ is the highest point in Θ_0. The LRT measures the "drop in altitude" between the global peak and the restricted peak. A large drop suggests the restriction is incompatible with the data.

10.5.2 Formal Definitions

Let the likelihood function be $L(\theta; x) = f(x; \theta)$. Consider

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta \setminus \Theta_0.$$

The **Likelihood Ratio** is

$$\lambda(x) = \frac{\sup_{\theta\in\Theta_0} L(\theta; x)}{\sup_{\theta\in\Theta} L(\theta; x)}.$$

The **Likelihood Ratio Test Statistic** is

$$\boxed{LR = -2\ln\lambda(x) = 2\left[\ell(\hat{\theta}) - \ell(\tilde{\theta})\right]},$$

where $\hat{\theta}$ is the unrestricted MLE and $\tilde{\theta}$ is the restricted MLE.

10.5.3 Non-negativity of the LR Statistic

Theorem: Under nested models ($\Theta_0 \subset \Theta$), $LR \geq 0$.

Proof:

1. Since $\Theta_0 \subset \Theta$, $\sup_{\Theta_0} L \leq \sup_{\Theta} L$.
2. Thus, $0 \leq \lambda(x) \leq 1$, so $-\infty < \ln\lambda \leq 0$.

3. Therefore, $LR = -2\ln\lambda \geq 0$. □

10.5.4 Wilks' Theorem

Theorem (Wilks): Under regularity conditions, if H_0 is true and θ_0 is in the interior of Θ,

$$\boxed{LR \xrightarrow{d} \chi^2_q},$$

where $q = \dim(\Theta) - \dim(\Theta_0)$ is the number of restrictions.

10.5.5 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Nested Models	$\Theta_0 \subset \Theta$ for hierarchical comparison	Non-nested models; Vuong's test required
Interiority of Parameter (Wilks)	Taylor expansion valid	Boundary test $H_0 : \sigma^2 = 0$; mixture distribution
Identifiability under Null	Unique likelihood maximum	Regime-switching models; parameter not identified

---

10.6 Likelihood Ratio Test — Applications

10.6.1 LRT for the Normal Mean (Unknown Variance)

Theorem: For $X_i \sim N(\mu, \sigma^2)$ i.i.d. with unknown variance, testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$,

$$\boxed{\lambda(X) = \left[1 + \frac{t^2}{n-1}\right]^{-n/2}},$$

where $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ is the usual t-statistic.

Proof:

1. The unrestricted MLEs are $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$.
2. Under H_0, $\mu = \mu_0$, so $\tilde{\sigma}^2 = \frac{1}{n}\sum(X_i - \mu_0)^2$.
3. The likelihood ratio is $\lambda(X) = (\hat{\sigma}^2/\tilde{\sigma}^2)^{n/2}$.
4. Decompose $\sum(X_i - \mu_0)^2 = \sum(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$.
5. Using $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ and $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$:

$$\lambda(X) = \left[1 + \frac{t^2}{n-1}\right]^{-n/2}.$$

□

10.6.2 LRT for the Exponential Parameter

Theorem: For $X_i \sim \text{Exp}(\lambda)$, testing $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$,

$$\boxed{\lambda(X) = (\lambda_0 \bar{X})^n e^{-n\lambda_0 \bar{X} + n}}.$$

Proof:

1. The likelihood is $L(\lambda) = \lambda^n e^{-\lambda \sum X_i}$.
2. The unrestricted MLE is $\hat{\lambda} = 1/\bar{X}$.
3. The likelihood ratio is
$$\lambda(X) = \frac{\lambda_0^n e^{-\lambda_0 \sum X_i}}{\hat{\lambda}^n e^{-\hat{\lambda} \sum X_i}}.$$
4. Substituting $\sum X_i = n/\hat{\lambda}$: $\lambda(X) = (\lambda_0 \bar{X})^n e^{-n\lambda_0 \bar{X} + n}$. □

10.6.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
i.i.d. Sampling	Joint likelihood factors into product	Autocorrelated data
Support Independent of Parameter (Wilks)	Asymptotic χ^2 distribution	Location-parameter Exponential; non-standard distribution

---

10.7 Nuisance Parameters in the LRT

10.7.1 Motivation

Nuisance parameters arise when the model requires several parameters, but the researcher is interested in only a subset. For example, when testing the mean μ of a normal population, the variance σ^2 is necessary but not of interest. The LRT elegantly handles this through **profiling**.

Geometric Interpretation: Decompose $\theta = (\psi, \lambda)$, where ψ is the parameter of interest and λ is nuisance. The **profiled likelihood** $L_p(\psi) = \sup_\lambda L(\psi, \lambda; x)$ "flattens" the mountain onto the axis of interest. The LRT compares the global peak with the peak on the hypersurface defined by H_0.

10.7.2 Formal Definitions

Decompose $\theta = (\psi, \lambda)$, where: - $\psi \in \Psi \subseteq \mathbb{R}^q$: parameters of interest. - $\lambda \in \Lambda \subseteq \mathbb{R}^{k-q}$: nuisance parameters.

Test $H_0 : \psi = \psi_0$ vs $H_1 : \psi \neq \psi_0$, with λ unrestricted.

The **Profiled Likelihood** is
$$\boxed{L_p(\psi) = \sup_{\lambda \in \Lambda} L(\psi, \lambda; x)}.$$

The **Profiled Likelihood Ratio Statistic** is
$$\boxed{\lambda_{LR} = \frac{L_p(\psi_0)}{\sup_{\psi \in \Psi} L_p(\psi)}}.$$

10.7.3 Wilks' Theorem Extension

Theorem: Under regularity conditions, if H_0 is true,
$$\boxed{LR = -2 \ln \lambda_{LR} \xrightarrow{d} \chi^2_q},$$
where $q = \dim(\psi)$ is the number of restrictions.

Proof: The profiled likelihood ratio is identical to the full LRT:

$$\lambda_{LR} = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}.$$

□

10.7.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Asymptotic Orthogonality	Estimates of ψ and λ asymptotically independent	Autocorrelated errors; non-block-diagonal information
Wilks Invariance	Asymptotic distribution depends only on q, not on λ_0	Unit root tests; Dickey-Fuller tables needed
Consistency of Nuisance Estimators	$\tilde{\lambda} \xrightarrow{p} \lambda_0$ under H_0	Incidental parameters problem (short panels)

10.8 The Neyman–Pearson Lemma

10.8.1 Motivation

The **Neyman–Pearson Lemma** solves the optimization problem of statistical tests. Given a fixed Type I Error (α), the researcher wants the test that **maximizes power** ($1 - \beta$). The lemma proves that, for simple hypotheses, the likelihood ratio test is the **Most Powerful (MP)** test.

Geometric Interpretation: In the sample space $\mathcal{X}$, under H_0 the data follow density $f_0(x)$; under H_1, density $f_1(x)$. Fixing α means choosing $R \subset \mathcal{X}$ with $\int_R f_0 = \alpha$. To maximize $\int_R f_1$, we should include points where $f_1(x)/f_0(x)$ is largest—the likelihood ratio.

10.8.2 Formal Statement

Consider testing simple $H_0 : f = f_0$ against simple $H_1 : f = f_1$. Let $\phi(x)$ be a test function:

$$\phi(x) = \begin{cases} 1, & \text{if } f_1(x) > k f_0(x), \\ 0, & \text{if } f_1(x) < k f_0(x), \end{cases}$$

for some $k \geq 0$, such that $\mathbb{E}_0[\phi(X)] = \alpha$.

Neyman–Pearson Lemma: Any test ϕ defined above is the **Most Powerful** test of level α. If ϕ' is any other test with $\mathbb{E}_0[\phi'(X)] \leq \alpha$, then

$$\mathbb{E}_1[\phi(X)] \geq \mathbb{E}_1[\phi'(X)].$$

10.8.3 Proof

1. Consider the power difference:

$$\Delta\pi = \int_{\mathcal{X}} (\phi(x) - \phi'(x)) f_1(x)\, dx.$$

2. By construction of ϕ, for all x:

$$(\phi(x) - \phi'(x))(f_1(x) - k f_0(x)) \geq 0.$$

3. Integrating:
$$\int (\phi - \phi')f_1\,dx \geq k \int (\phi - \phi')f_0\,dx.$$

4. The right-hand side is $k(\alpha - \mathbb{E}_0[\phi']) \geq 0$.

5. Therefore, $\mathbb{E}_1[\phi] - \mathbb{E}_1[\phi'] \geq 0$, so $\pi \geq \pi'$. $\square$

10.8.4 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Simple Hypotheses	Both H_0 and H_1 specify a single distribution	Composite $H_1 : \mu > \mu_0$; UMP may not exist
Existence of Densities	Likelihood ratio computable pointwise	Singular distributions; trivial test
Fixed Significance Level	Constraint for power maximization	No α constraint; "always reject" trivial test

---

10.9 Most Powerful and Uniformly Most Powerful Tests

10.9.1 Formal Definitions

A test ϕ^* of size α is **Most Powerful (MP)** if, for any other test ϕ of size $\alpha' \leq \alpha$,

$$\mathbb{E}_1[\phi^*(X)] \geq \mathbb{E}_1[\phi(X)].$$

A test ϕ^* is **Uniformly Most Powerful (UMP)** if, for any other test ϕ of size α and for **every** $\theta \in \Theta_1$,

$$\boxed{\pi^*(\theta) = \mathbb{E}_\theta[\phi^*(X)] \geq \mathbb{E}_\theta[\phi(X)] = \pi(\theta), \quad \forall \theta \in \Theta_1}.$$

10.9.2 Monotone Likelihood Ratio

Definition: A family of densities $\{f(x;\theta)\}$ has **Monotone Likelihood Ratio (MLR)** in a statistic $T(x)$ if, for $\theta_2 > \theta_1$, the ratio $f(x;\theta_2)/f(x;\theta_1)$ is non-decreasing in $T(x)$.

Theorem: If the family has MLR in $T(x)$, then the UMP test for $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ rejects when $T(x) > c$.

10.9.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Monotone Likelihood Ratio	Evidence grows consistently with $T(x)$	Two-sided alternatives; UMP does not exist
One-Sided Alternative	Direction of deviation is consistent	Two-sided $H_1 : \theta \neq \theta_0$

10.10 The p-Value

10.10.1 Motivation

The **p-value** solves the problem of **arbitrariness in the choice of** α. The classical Neyman–Pearson approach requires fixing α before observing the data, creating a rigid binary decision. The p-value transforms this discrete decision into a continuous measure of evidence.

Geometric Interpretation: The p-value is the **area under the curve** (probability mass) from the observed test statistic t_{obs} toward the tail defined by the alternative. It maps any test statistic to $[0, 1]$, representing the "rarity" of the observed result under H_0.

10.10.2 Formal Definition

Let $T(X)$ be a test statistic such that larger values provide evidence against H_0. Let $G_0(t) = \mathbb{P}(T \leq t \mid H_0)$ be the CDF under the null. The **p-value** is

$$\boxed{p(x) = \mathbb{P}(T(X) \geq T(x) \mid H_0)}.$$

Equivalently, the p-value is the **smallest significance level** at which H_0 would be rejected:

$$p = \inf\{\alpha \in (0, 1) : x \in R_\alpha\},$$

where R_α is the critical region of size α.

10.10.3 Distribution of the p-Value

Theorem: If H_0 is true and G_0 is continuous, then

$$\boxed{P \sim U(0, 1)}.$$

Proof (one-sided upper test):

1. $P = 1 - G_0(T)$.
2. For $u \in [0, 1]$,
$$F_P(u) = \mathbb{P}(P \leq u \mid H_0) = \mathbb{P}(1 - G_0(T) \leq u).$$
3. Rearranging: $F_P(u) = \mathbb{P}(T \geq G_0^{-1}(1 - u))$.
4. By complement: $F_P(u) = 1 - G_0(G_0^{-1}(1 - u)) = u$.
5. Therefore, $F_P(u) = u$, the CDF of $U(0, 1)$. $\square$

10.10.4 Equivalence of p-Value and Critical Region

Theorem: The decision rule $p < \alpha$ is equivalent to $T(x) > c_\alpha$.

Proof:

1. Reject if $T(x) > c_\alpha = G_0^{-1}(1 - \alpha)$.
2. Apply G_0: $G_0(T(x)) > 1 - \alpha$.
3. Rearrange: $1 - G_0(T(x)) < \alpha$.
4. Since $p = 1 - G_0(T(x))$: $p < \alpha$. $\square$

10.10.5 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Knowledge of G_0	Tail probability computable	Unknown null distribution
Clear Alternative Direction	Determines which tail is "extreme"	Ambiguous H_1; p-value undefined
Continuity of T	p-value $\sim U(0, 1)$	Discrete Binomial; staircase distribution

---

10.11 Interpretation of the p-Value

10.11.1 Formal Interpretation

The p-value is the probability, assuming H_0 is true, of obtaining a test statistic as extreme as or more extreme than the one actually observed.

Key Properties: 1. If H_0 is true, $p \sim U(0,1)$ (for continuous statistics). 2. For any α, $\mathbb{P}(p \leq \alpha \mid H_0) = \alpha$. 3. The p-value is **not** the probability that H_0 is true. 4. The p-value is **not** the probability of a Type I Error.

10.11.2 Common Misinterpretations

Misinterpretation	Correction
"The p-value is the probability that H_0 is true."	The p-value is computed assuming H_0 is true; it does not quantify the probability of the hypothesis itself.
"A p-value of 0.05 means there is a 5% chance of a Type I Error."	The Type I Error rate is fixed at α before the experiment, not derived from the observed p-value.
"Smaller p-values always mean larger effects."	The p-value depends on sample size; a small p-value can arise from a tiny effect with a large sample.
"If $p > 0.05$, H_0 is true."	Failure to reject H_0 does not imply H_0 is true; it only indicates insufficient evidence to reject.

10.11.3 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Correct Specification under H_0	p-value scale is reliable	Small sample without normality; $p = 0.04$ is an artifact
Clear Alternative Direction	Determines which tail to integrate	One-sided vs two-sided; different p-values
p-Value as a Random Variable	Recognized as a statistic, not a parameter	Interpreting $p = 0.03$ as "3% chance H_0 is true"

10.12 Summary of Key Results

Concept	Definition	Key Formula
Null Hypothesis	Statement about θ	$H_0 : \theta \in \Theta_0$
Alternative Hypothesis	Statement about θ	$H_1 : \theta \in \Theta_1$
Test Statistic	Function of the sample	$T : \mathcal{X} \to \mathbb{R}$
Critical Region	Values leading to rejection	$R = \{t : t > c\}$
Critical Value	Boundary of rejection	$c = G_0^{-1}(1-\alpha)$
Type I Error	Reject H_0 when true	$\alpha = \mathbb{P}(T \in R \mid H_0)$
Type II Error	Fail to reject when false	$\beta = \mathbb{P}(T \notin R \mid H_1)$
Power	Probability of correct rejection	$\pi(\theta) = 1 - \beta(\theta)$
Likelihood Ratio	Ratio of maximized likelihoods	$\lambda = \frac{\sup_{\Theta_0} L}{\sup_{\Theta} L}$
LR Statistic	Log of likelihood ratio	$LR = -2 \ln \lambda$
Wilks' Theorem	Asymptotic distribution	$LR \xrightarrow{d} \chi^2_q$
Neyman–Pearson Lemma	MP test for simple hypotheses	Reject if $f_1(x)/f_0(x) > k$
p-Value	Tail probability under H_0	$p = 1 - G_0(T(x))$
p-Value Interpretation	Smallest α for rejection	$p = \inf\{\alpha : x \in R_\alpha\}$

Chapter 11: Confidence Intervals

11.1 Construction via Pivotal Quantities

11.1.1 Definition of Pivotal Quantity

11.1.1.1 Motivation The fundamental problem that confidence interval construction solves is the **quantification of uncertainty** inherent in a point estimate. Although an estimator $\hat{\theta}$ provides the "most likely" value for a fixed population parameter θ, it will rarely coincide exactly with the true value due to sampling variability. The pivotal method provides a systematic technique for transforming a point estimate into a random region that, with a pre-specified probability, "traps" the target parameter.

Geometric Interpretation: In one dimension, the confidence interval should not be interpreted as the probability that the fixed parameter θ "falls" within an interval. Instead, imagine θ as a fixed point on the real axis. Each possible sample generates a different interval (a line segment). If we set a confidence level of 95%, then if we repeated the experiment infinitely many times, 95% of these randomly generated segments would cover the fixed point θ.

11.1.1.2 Fundamental Hypotheses

Hypothesis	Role	Counterex
i.i.d. Sampling	Ensures $\text{Var}(\bar{X}) = \sigma^2/n$	Autocorre below nor
Existence of a Pivotal Quantity	Distribution of $Q(\mathbf{X}, \theta)$ independent of θ	Distributi no univers

---

11.1.1.3 Formal Definition Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $F(x; \theta)$, where $\theta \in \Theta$ is the parameter of interest.

Definition (Pivotal Quantity): A function $Q(\mathbf{X}, \theta)$ is called a **pivotal quantity** (or simply a *pivot*) if its sampling distribution is the same for all values of $\theta \in \Theta$.

Theorem (Construction of the CI): If $Q(\mathbf{X}, \theta)$ is a pivotal quantity with known CDF $G(q)$, then for any $\alpha \in (0, 1)$, there exist critical values $q_{\alpha/2}$ and $q_{1-\alpha/2}$ such that

$$\boxed{\mathbb{P}(q_{\alpha/2} \leq Q(\mathbf{X}, \theta) \leq q_{1-\alpha/2}) = 1 - \alpha}.$$

If Q is monotone in θ, the interval can be inverted to obtain $\hat{C} = [L(\mathbf{X}), U(\mathbf{X})]$ such that $\mathbb{P}(\theta \in \hat{C}) = 1 - \alpha$.

11.1.1.4 Derivation for Normal Mean with Known σ Consider the classical case where $X_i \sim N(\mu, \sigma^2)$ with σ^2 known.

1. Identification of the Estimator:

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

2. Distribution of the Estimator:

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

3. Definition of the Pivotal Quantity:

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

Since the distribution $N(0, 1)$ has no unknown parameters, Z is a **pivotal quantity**.

4. Establishing the Coverage Probability: For confidence level $1 - \alpha$,

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha.$$

5. Inversion of the Pivot: Substituting the definition of Z:

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

Multiplying by $\sigma/\sqrt{n}$:

$$-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Isolating μ:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Conclusion: The confidence interval is

$$\boxed{CI(\mu; 1-\alpha) = \left[\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]}.$$

The hypothesis that $Z \sim N(0,1)$ is **independent of** μ was indispensable; without it, the critical values would depend on the very parameter we are trying to estimate.

11.1.2 Examples: Mean with Known and Unknown Variance

11.1.2.1 Motivation The technical objective of these examples is to demonstrate how **studentization** (adjusting the scale of a centered statistic by a dispersion estimator) alters the geometry of the resulting interval. When σ is known, the uncertainty comes solely from the position of $\bar{X}$. Geometrically, the interval length is fixed for a given sample size.

When σ is unknown, substituting σ with S introduces a second source of random variability. This results in intervals of variable length. To compensate for the additional uncertainty about the scale, the reference distribution has heavier tails, requiring larger quantiles to maintain the same coverage level.

11.1.2.2 Fundamental Hypotheses

Hypothesis	Role	Counterex
Population Normality	Ensures $\bar{X} \perp S^2$; exact t-distribution	Skewed p quantiles
Sample Independence (SRS)	$\mathrm{Var}(\bar{X}) = \sigma^2/n$	Positive a underesti

---

11.1.2.3 Case A: Mean with Known σ^2 **Formal Statement:** Let $X_i \sim N(\mu, \sigma^2)$ i.i.d. with σ^2 known.

1. Pivotal Quantity:

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

2. Acceptance Region:

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1-\alpha.$$

3. Inversion:

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Result:

$$\boxed{CI(\mu; 1-\alpha) = \left[\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].}$$

11.1.2.4 Case B: Mean with Unknown σ^2 **Theorem (Simplified Cochran):** If $X_i \sim N(\mu, \sigma^2)$ i.i.d., then: (i) $\bar{X}$ and S^2 are independent. (ii) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

1. Construction of the Student's t Statistic:

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Since the distribution depends only on $n-1$ and not on μ or σ, T is a **pivotal quantity**.

2. Establishing the Probability:

$$\mathbb{P}(-t_{1-\alpha/2,n-1} \leq T \leq t_{1-\alpha/2,n-1}) = 1-\alpha.$$

3. Inversion:

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

Result:

$$\boxed{CI(\mu; 1-\alpha) = \left[\bar{X} \pm t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right].}$$

The hypothesis of **normality** was indispensable in the derivation of T; without it, $\bar{X}$ and S^2 would not necessarily be independent, preventing the statistic from following an exact Student's t-distribution.

11.2 Confidence Intervals for the Mean

11.2.1 With Known Variance

11.2.1.1 Motivation The problem of estimating the population mean μ with known variance σ^2 is the fundamental scenario of statistical inference. The objective is to move from a point estimate $\bar{X}$ to an interval estimate that provides a margin of error based on the inherent variability of the process.

Geometric Interpretation: The confidence interval is a line segment centered at $\bar{x}$. Since σ^2 is known, the width of this interval is **constant** for a given n and confidence level γ. The interval is random because its center fluctuates between samples, but its width is deterministic.

11.2.1.2 Fundamental Hypotheses

Hypothesis	Role	Counterex
SRS (i.i.d.)	$\text{Var}(\bar{X}) = \sigma^2/n$	Cluster sa larger
Normality or Large n	Sampling distribution of $\bar{X}$ is normal	Heavy-tai CLT fails
Known σ^2	Allows use of Z-distribution	Unknown underesti

11.2.1.3 Formal Derivation **1. Pivotal Quantity:**

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

2. Probability Statement:

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha.$$

3. Inversion:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Result:

$$\boxed{CI(\mu; 1-\alpha) = \left[\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].}$$

11.2.2 With Unknown Variance (Student's t)

11.2.2.1 Motivation In practice, σ^2 is almost always unknown. To overcome this, we substitute σ with its unbiased estimator S. This process is called **studentization**.

Geometry and Uncertainty: Unlike the known variance case, the interval width is a **random variable**, as it depends on S. The Student's t-distribution has **heavier tails** than the normal, reflecting additional uncertainty: larger quantiles are needed to guarantee the same coverage probability $1 - \alpha$.

11.2.2.2 Fundamental Hypotheses

Hypothesis	Role	Counterex
Population Normality	Guarantees $\bar{X} \perp S^2$; exact t-distribution	Cauchy; v
SRS (i.i.d.)	Sum of squared deviations follows χ^2	Spatially underesti

---

11.2.2.3 Formal Derivation **1. Unbiased Sample Variance:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

2. Chi-square Distribution:

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

3. Student's t Statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Since the distribution does not depend on unknown parameters, T is a **pivot**.

4. Probability Statement:

$$\mathbb{P}(-t_{1-\alpha/2,n-1} \leq T \leq t_{1-\alpha/2,n-1}) = 1 - \alpha.$$

5. Inversion:

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

Result:

$$\boxed{CI(\mu; 1-\alpha) = \left[\bar{X} \pm t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right].}$$

The hypothesis of **normality** was indispensable for applying Cochran's Theorem, ensuring that the ratio results in an exact t-distribution.

11.3 Confidence Intervals for the Variance

11.3.1 Using the χ^2 Distribution

11.3.1.1 Motivation Inference about the population variance σ^2 is fundamental in contexts where the **stability or precision** of a process is more critical than its central location. While intervals for the mean deal with "where" the data are, the interval for the variance quantifies "how spread out" they are.

Geometric Interpretation: Unlike the mean construction, which relies on symmetry, the variance construction uses the **Chi-square (χ^2)** distribution, defined on $\mathbb{R}^+$ with **right skewness**. The resulting interval for σ^2 is **asymmetric** around S^2, and naturally respects the lower bound of zero.

11.3.1.2 Fundamental Hypotheses

Hypothesis	Role	Counterex
Population Normality	$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$	Heavy tai than nom
SRS (i.i.d.)	Sum of squares follows χ^2 definition	Positive s underesti

---

11.3.1.3 Formal Derivation **1. Sample Variance:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

2. Chi-square Distribution (Cochran's Theorem):

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Proof Sketch:

1. Define $Z_i = (X_i - \mu)/\sigma \sim N(0,1)$.
2. Decompose:
$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$
3. Divide by σ^2:
$$\sum_{i=1}^{n} Z_i^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$
4. The left side is χ^2_n, the second term is χ^2_1. By Cochran's Theorem, the first term is χ^2_{n-1} and independent of $\bar{X}$.

3. Probability Statement: Let $q_{\alpha/2,n-1}$ and $q_{1-\alpha/2,n-1}$ be the quantiles of χ^2_{n-1}:

$$\mathbb{P}\left(q_{\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq q_{1-\alpha/2,n-1}\right) = 1 - \alpha.$$

4. Inversion:

$$\frac{(n-1)S^2}{q_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{q_{\alpha/2,n-1}}.$$

Result:

$$\boxed{CI(\sigma^2; 1-\alpha) = \left[\frac{(n-1)S^2}{q_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{q_{\alpha/2,n-1}}\right].}$$

The hypothesis of **normality** was indispensable for step 4; without it, the sum of squared deviations would not follow an exact Chi-square distribution.

11.4 One-Sided Confidence Intervals

11.4.1 Motivation

In many applications, the researcher's interest is not in "surrounding" the parameter on both sides, but in establishing a **safety limit** or **critical threshold**. For example: ensuring average strength is *at least* a certain value (lower bound), or ensuring maximum loss does not exceed a threshold (upper bound).

Geometric Interpretation: Unlike the two-sided interval, which distributes α equally across two tails ($\alpha/2$ each), the one-sided interval concentrates the entire error mass in a single tail. This transforms the interval into a **ray** extending to infinity. For the same confidence level, the bound of a one-sided interval is closer to the point estimate than the corresponding bound of a two-sided interval.

11.4.2 Fundamental Hypotheses

Hypothesis	Role	Counterexample
Known Sampling Distribution	CDF of pivot is known	Unknown distribution; insufficient n for CLT
Monotonicity of Pivot	Allows inversion to isolate θ	Quadratic pivot; inversion yields disjoint sets

---

11.4.3 Formal Derivation

Consider the pivot $Z = \frac{\hat{\theta}-\theta}{s(\hat{\theta})} \sim N(0,1)$. Let z_p be the p-quantile such that $\mathbb{P}(Z \leq z_p) = p$.

11.4.3.1 Lower One-Sided Interval **Objective:** Find $L(\mathbf{X})$ such that $\mathbb{P}(\theta \geq L(\mathbf{X})) = 1-\alpha$.

1. Probability Statement:

$$\mathbb{P}(Z \leq z_{1-\alpha}) = 1-\alpha.$$

2. Substitution:

$$\mathbb{P}\left(\frac{\hat{\theta}-\theta}{s(\hat{\theta})} \leq z_{1-\alpha}\right) = 1-\alpha.$$

3. Algebraic Manipulation:

$$\hat{\theta} - \theta \leq z_{1-\alpha} s(\hat{\theta}).$$

$$-\theta \leq z_{1-\alpha} s(\hat{\theta}) - \hat{\theta}.$$

$$\theta \geq \hat{\theta} - z_{1-\alpha} s(\hat{\theta}).$$

Result:

$$\boxed{CI_{\text{lower}}(\theta; 1-\alpha) = \left[\hat{\theta} - z_{1-\alpha} s(\hat{\theta}), \infty\right).}$$

11.4.3.2 Upper One-Sided Interval **Objective:** Find $U(\mathbf{X})$ such that $\mathbb{P}(\theta \leq U(\mathbf{X})) = 1 - \alpha$.

1. Probability Statement:

$$\mathbb{P}(Z \geq -z_{1-\alpha}) = 1 - \alpha.$$

(by symmetry: $z_\alpha = -z_{1-\alpha}$)

2. Substitution:

$$\mathbb{P}\left(\frac{\hat{\theta} - \theta}{s(\hat{\theta})} \geq -z_{1-\alpha}\right) = 1 - \alpha.$$

3. Algebraic Manipulation:

$$\hat{\theta} - \theta \geq -z_{1-\alpha} s(\hat{\theta}).$$

$$-\theta \geq -z_{1-\alpha} s(\hat{\theta}) - \hat{\theta}.$$

$$\theta \leq \hat{\theta} + z_{1-\alpha} s(\hat{\theta}).$$

Result:

$$\boxed{CI_{\text{upper}}(\theta; 1-\alpha) = \left(-\infty, \hat{\theta} + z_{1-\alpha} s(\hat{\theta})\right]}.$$

Final Note: The hypothesis of **symmetry** (Normal or t) simplifies the quantiles, but the construction via test inversion works for asymmetric distributions (like χ^2) using the specific quantiles of that distribution. The hypothesis of **known variance** or **large sample size** was indispensable for using z quantiles; otherwise, quantiles from the t-distribution with $n - k$ degrees of freedom would be required.

11.5 Summary of Key Results

Concept	Definition	Key Formula
Pivotal Quantity	Function whose distribution is independent of θ	$Q(\mathbf{X}, \theta) \sim G$ (known)
Confidence Interval	Random interval covering θ with probability $1 - \alpha$	$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha$
CI for Mean (Known σ)	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$
CI for Mean (Unknown σ)	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$	$\bar{X} \pm t_{1-\alpha/2, n-1} \frac{S}{\sqrt{n}}$
CI for Variance (σ^2)	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$	$\left[\frac{(n-1)S^2}{q_{1-\alpha/2}}, \frac{(n-1)S^2}{q_{\alpha/2}}\right]$
Lower One-Sided CI	$\theta \geq \hat{\theta} - z_{1-\alpha} s(\hat{\theta})$	$[\hat{\theta} - z_{1-\alpha} s(\hat{\theta}), \infty)$
Upper One-Sided CI	$\theta \leq \hat{\theta} + z_{1-\alpha} s(\hat{\theta})$	$(-\infty, \hat{\theta} + z_{1-\alpha} s(\hat{\theta})]$
Width of CI	Depends on n, α, and $s(\hat{\theta})$	Width $= 2 \cdot z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
Coverage Probability	$\mathbb{P}(\theta \in CI)$	$1 - \alpha$ (nominal)
Studentization	Replacing σ with S	T-statistic follows t_{n-1}
Cochran's Theorem	Independence of $\bar{X}$ and S^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
