Probability and Statistical Inference

Tianqi Zhang

Emory University

Apr 17th 2025

Contents

Preface

Welcome to my study notes on Probability and Statistical Inference, where we delve into foundational concepts in probability theory through the lens of measure theory. The notes aim to provide a structured journey that bridges rigorous mathematical principles with their practical applications. These notes are structured for readers who have a solid understanding of basic real analysis, including:

- Sequence and function convergence and limit
- Point-set topology
- Integration theory and Riemann integral

By following this text, you will explore topics ranging from fundamental randomness and sigma-algebras to Lebesgue integration and advanced probability concepts such as stochastic convergence and density transformations. The material is organized to balance theoretical depth with illustrative examples, ensuring clarity and progression in understanding.

The primary goal of this document is to serve as both a companion for self-study and a reference for further research in probability theory, statistics, and related mathematical fields.

I hope you find this document insightful as you navigate the beauty and rigor of probability and measure theory.

Preliminary definitions

Definition 0.0.1 (Set convergence).

For any general sequence of sets $A_n \to A$ if

$$\mathbb{I}_{A_n}(\omega) \to \mathbb{I}_A(\omega) \quad \text{for all } \omega \in \Omega$$

1 Randomness

1.1 Randomness: A Model of Empirical Observations

Definition 1.1.1 (Latent Space Ω).
We denote the latent space Ω to be the set of all possible outcomes.
Random: A model of an empirically observed property of the world.

Definition 1.1.2 (Random Variable). The random variable X maps each $\omega \in \Omega$ to $\mathbb{R}^d$

Example 1.1.3 Coin toss
The random variable $X : \{\text{Head, Tail}\} \to \mathbb{R}$ is defined to be:

$$X(\text{Head}) = 1, \quad X(\text{Tail}) = 0$$

If we repeat the same experiment under the same conditions, an event $A \subseteq \Omega$ will occur in some experiments but not in others.

1.2 Sampling Frequency

If we conduct n experiments, event A occurs exactly n_A times. Then the sampling frequency of A is:

$$f_n(A) = \frac{n_A}{n}.$$

Example 1.2.1 Coin toss
$\{3 \text{ heads, } 97 \text{ tails}\}$ implies in 100 random experiments with a fair coin:

$$f_{100}(3 \text{ heads, } 97 \text{ tails}) = \frac{3}{100}.$$

1.3 Two Philosophies of Randomness

What happens as $n \to \infty$?

1.3.1 Frequentist Inference

- Stability at large scale: volatility of fluctuations of $f_n(A)$ tends to decrease, where $n \to \infty$.

Theorem 1.3.1 Frequentists' idea
Existence of population probability $P(A) \in [0, 1]$, that is not random, where

$$P(A) = \lim_{n \to \infty} f_n(A)$$

.

However, this limit cannot be deterministic. For example, bad events like:

$$B_n = \{|f_n(A) - P(A)| \geq \varepsilon\}$$

may still occur for large n, which paves the way for the laws of large numbers (LLN) to control B_n.

Remark 1.1 *For all* $\varepsilon > 0$, $\Pr(B_n$ *happens*$) \to 0$ *as* $n \to \infty$, *where:*

$$B_{n,\varepsilon} = \{|f_n(A) - P(A)| \geq \varepsilon\}.$$

1.3.2 Probability Theory vs. Statistical Inference

Remark 1.2 *Probability Theory: For given $P(A)$, compute the probability that a future series of n events lies in an interval:*

$$f_n(A) \in [P(A) - \varepsilon, P(A) + \varepsilon].$$

Remark 1.3 *Statistical Inference: For given statistical evidence,*

1. *A point estimate for $P(A)$:*
$$\lim_{n \to \infty} f_n(A) = P(A)$$
2. *A* ***confidence interval estimate*** *for $P(A)$:*
$$CI_n = [f_n(A) - Z_n, f_n(A) + Z_n]$$
Such that $\lim_{n \to \infty} Pr\{P(A) \subset CI_n\} \geq 1 - \alpha$ *(commonly* $\alpha = 0.05$*)*

So the two remarks are the inverse problems of each other.

Definition 1.3.2 (Estimator). An **estimator** is a computational rule with given data (a function, or later defined as a random variable of data).

Independent experiments help reduce fluctuations by leveraging concentration of measure results.

1.3.3 Bayesian Estimation

Theorem 1.3.3 Bayesian Idea
The unknown $p = P(A)$ is itself random, and Statistical data reduces uncertainty of the parameter (information update).

- **Prior:** What we believe about $P(A)$ before gathering data.
- **Posterior:** Updated beliefs after gathering data.

The Bernstein-von Mises Theorem might offer some insights reconciling the two schools of thoughts when n is sufficiently large.

Theorem 1.3.4 Bernstein-von Mises Theorem
The posterior distribution is independent of the prior distribution (under some conditions) once the amount of information supplied by a sample of data is large enough

2 Sampling

2.1 Sample Space and Probability Models

Definition 2.1.1 (Sample Space).
The specification of a probability model requires:

1. A sample space, Ω: The set of all possible outcomes in the problem.

2. A probability assignment for these outcomes (subsets of Ω).

Consider the following examples to help understanding the construction of probability models:

Example 2.1.2 Coin Tossing

$$\Omega = \{\mathrm{H}, \mathrm{T}\}.$$

- If assuming the coin is balanced, we state $P(\mathrm{H}) = P(\mathrm{T}) = \frac{1}{2}$.

- This assumption can be extended to all subsets of Ω.

- Assumption can be justified empirically using the **Law of Large Numbers (LLN)**:

  $$P(A) = \operatorname*{plim}_{n\to\infty} f_n(A).$$

  For a proper definition of "plim"

To generalize the above results:

Definition 2.1.3 (Uniform probability measure/distribution).

- $\Omega = \{\omega_1, \ldots, \omega_N\}$ or countable

- **Uniform probability measure (distribution)**: For any $A \subseteq \Omega$,

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

  and define

  $$P(\omega_i) = \frac{1}{N}, \forall i \in \{1, 2, \ldots, N\}$$

- **Warning on abuse of notation:** Note that this is not a proper probability measure, as it assigns probabilities to individual elements, not subsets.

  $$P(\omega_i) \equiv P(\{\omega_i\})$$

  even though $\omega_i \in \Omega$, $\{\omega_i\} \subseteq \Omega$, and $\{\omega_i\} \in \mathscr{P}(\Omega)$ as the power set.

Therefore, we have the above function P defined to be $\mathscr{P}(\Omega) \to [0,1]$

Example 2.1.4 Fair Die

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

- Let $A =$ "even numbers" $= \{2,4,6\}$. Then:

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

Remark 2.1 *Conditioning and Specifying Ω is crucial*

Example 2.1.5 Twin paradox example

- **Assumptions:**
  - (i) Gender of newborn: $P(\text{Girl}) = P(\text{Boy}) = \frac{1}{2}$.
  - (ii) Gender of one child is independent of the gender of the other child.
- Known: A family has two children, one of whom is a girl.
- Question: What is P(both children are girls)?
  - Case (i): If we know the gender of the first child, then Ω considers only the second child

    $$\Omega_i = \{\text{G}, \text{B}\}.$$

    Then:

    $$P(\text{G}) = \frac{1}{2}.$$

  - Case (ii): If we know at least one child is a girl:

    $$\Omega_{ii} = \{\text{GG}, \text{GB}, \text{BG}\}.$$

Then:

$$P(\text{GG}) = \frac{1}{3}.$$

2.2 Bernoulli Trials and Binomial Distribution

Example 2.2.1 Bernoulli Distribution

- Suppose that we have a sequence of n independent experiments

$$P(\text{Success}) = p,\ \ P(\text{Failure}) = 1 - p$$

- We have the sample space:

$$\Omega = \{0, 1\}^n \quad \text{(Cartesian product of the simple } \{0, 1\}\text{)}.$$

- A typical element (draw) $\omega \in \Omega$ is $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ as a sequence.
- Define $S_n : \Omega \to \mathbb{R}$, $S_n(\omega) = \sum_{i=1}^{n} \omega_i$ to be the count of success.

Claim 2.2.2:
Then the **independent experiment** with its probability measure:

$$P(\omega) = p^{S_n(\omega)}(1 - p)^{n - S_n(\omega)}$$

Remark 2.2 *S_n is a* ***random variable****, a function from Ω endowed with the probability measure P and maps outcomes to $\mathcal{S} \equiv \{0, 1, \ldots, n\}$.*

Since the original Bernoulli trial is simple enough to be binary: $\Omega = \{0, 1\}$, the count of success can be reduced into the simple sum in the sequence. Therefore, we can then define a new probability measure (distribution) P^{S_n} on S such that:

Theorem 2.2.3 Push-Forward Measure
We can then define a new probability measure (distribution) P^{S_n} on S such that

$$\forall x \in S, P(x) \equiv \sum_{\omega \in S_n^{-1}(x)} P(\omega) = P\big[S_n^{-1}(x)\big] \equiv P^{S_n}(x)$$

Where

$$S_n^{-1}(A) \equiv \{\omega \in \Omega, S_n(\omega) \in A\} \quad \text{is the pre-image}$$

For computation purposes:

$$\forall x \in \mathcal{S},\ P^{S_n}(x) = \big|S_n^{-1}(x)\big| p^x (1-p)^{n-x}$$

Intuitively speaking, the pre-image $S_n^{-1}(x)$ indicates all possible sequences ω such that they contain exactly x counts of success. We give the calculation as follows

Definition 2.2.4 (Binomial number).

$$\big|S_n^{-1}(x)\big| = \binom{n}{x} = \frac{n!}{x!(n-x)!} = \binom{n}{n-x}$$

To be the number of ways to pick x elements from n elements without order.

Definition 2.2.5 (Binomial Distribution).
Binomial Distribution: $Binom(n, p)$ is the push-forward probability measure of the Bernoulli trial with n experiments and success probability p via the random variable S_n which counts success.

$$\forall x \in \mathcal{S}, P^{S_n}(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

and

$$\sum_{n=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$$

We say the **random variable S_n follows the binomial distribution**, $S_n \sim Binom(n, p)$.

3 Discrete Probability Measures

3.1 Discrete Probability Measures

We start with discrete, countable latent space Ω.

Definition 3.1.1 (Discrete Probability Measure). A discrete probability measure on sample space Ω, finite or countable, is a sequence of $\{p_\omega\}_{\omega \in \Omega}$ of non-negative real numbers such that:

1. $p_\omega \geq 0$, $\forall \omega \in \Omega$,
2. $\sum_{\omega \in \Omega} p_\omega = 1$

A general definition that works not only for finite Ω but also for countable Ω since it allows in both cases to compute for any random event $A \subseteq \Omega$.

$$P(A) = \sum_{\omega \in A} P_\omega$$

Definition 3.1.2 (Measure/Distribution).
A measure P on Ω is a mapping from the power set of the latent space Ω.

$$P : \mathscr{P}(\Omega) \to [0, \infty]$$

such that the following two axioms are satisfied:

1. **Non-negativity:** $\forall A \subseteq \Omega, P(A) \geq 0$
2. **Countable additivity:** (Or sigma additivity) For disjoint $A_n \subset \Omega$,

$$P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_k P(A_n)$$

3. **Empty set measurability:** $P(\varnothing) = 0$.

Note that for P to be a probability measure: $P(A) \in [0, 1]$ $\forall A \in \mathscr{P}(\Omega)$, with $P(\Omega) = 1$.

Theorem 3.1.3 Komolgrov Axioms
$P : \mathscr{P}(\Omega) \to [0, 1]$ is a probability measure if the above conditions are true.

Remark 3.1 *inclusion of infinity: In most measure-theoretic contexts, it is permissible for certain subsets $\mathcal{A} \subset \Omega$ to have infinite measure. Consequently, the codomain of a measure is typically extended to include infinity. This is commonly represented as the set of positive real numbers together with infinity, denoted by $[0, \infty) \cup \{\infty\}$. For simplicity, this notation is often abbreviated as $[0, \infty]$.*

Example 3.1.4 Common Measures

- **Counting Measure:** $\mu(A) = |A|$.
- **Dirac Measure at $p \in \Omega$:**

$$\delta_p(A) = \begin{cases} 1, & \text{if } p \in A, \\ 0, & \text{otherwise.} \end{cases}$$

- **Lebesgue Measure on $\mathbb{R}$:** For simple intervals $[a, b) \subset \mathbb{R}$ with $a \leq b$. Lebesgue measure μ is defined to be $b - a$.

Proposition 3.1.5 Properties of Komolgrov axioms:

1. $P(\overline{A}) = 1 - P(A)$ where $\overline{A} \equiv \Omega \backslash A$.
2. Define $A + B$ as the disjoint union of A, B, i.e., $A \cap B = \varnothing$, then $P(A + B) = P(A) + P(B)$.
3. If $B \subset A$, define $A - B \equiv A \cap \overline{B}$ then $P(A - B) = P(A) - P(B)$.
4. Partition Ω into disjoint $\{H_i\}_{i \in \mathbb{N}}$, i.e. $H_i \cap H_j = \varnothing \ \forall i, j$ and $\bigcup_i H_i = \Omega$. Then $\forall A \subset \Omega, P(A) = \sum_i P(A \cap H_i)$.

Corollary 3.1.6 Monotonicity:
If $A \subseteq B$, then $0 \leq P(A) \leq P(B) \leq 1$.

Corollary 3.1.7 Sylvester Formula:
For any collection of subsets $\{A_i\}$ of Ω

$$P\left(\bigcup_{i=1}^{n}\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right)$$

Or a more intuitive version, for any $A, B \subseteq \Omega$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Remark 3.2 *Banach-Tarski Paradox: If Ω is countable, we can assign a finite measure to all subsets of Ω satisfying the Kolmogorov axioms. However, if Ω is uncountable, these axioms can lead to paradoxes.*

To address these issues, we may need to restrict our measure $P : \mathscr{P}(\Omega) \to [0, \infty]$ to a carefully chosen collection of subsets $\mathcal{F} \subset \Omega$. This restriction sets the foundation for further discussion on the role of σ-algebras in measure theory.

3.2 Results and Properties

Proposition 3.2.1 Boole's Inequality:
For $A_i \subseteq \Omega$, not necessarily disjoint:

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i).$$

Proof: A.1.1
Proposition 3.2.2 Bonferroni's Inequality:
For $A_i \subseteq \Omega$:

$$P\left(\bigcap_i A_i\right) \geq \sum_i P(A_i) - (n-1).$$

This result is useful for multiple hypothesis testing.
Proof: A.1.2
Proposition 3.2.3 De Morgan's Laws:

$$P\left(\bigcap_i A_i\right) = P\left(\bigcup_i \overline{A_i}\right).$$

Some notations for limits of sets:

- **Increasing Sequence:** (A_n) is called an increasing sequence if $(A_n \subseteq A_{n+1})$, and

$$\lim_{n\to\infty} \uparrow A_n \equiv \bigcup_{k=1}^{\infty} A_n$$

- **Decreasing Sequence:** (B_n) is called an increasing sequence if $(B_n \supseteq B_{n+1})$, and

$$\lim_{n\to\infty} \downarrow B_n \equiv \bigcap_{k=1}^{\infty} B_n$$

Proposition 3.2.4 Continuity of Measures:

The continuity of measures is preserved under increasing and decreasing set limits.

Let (A_n) be an increasing sequence of sets:
Increasing continuity:

$$\lim_{n\to\infty} P(A_n) = P\Big(\lim_{n\to\infty} \uparrow A_n\Big) \equiv P\left(\bigcup_{n=1}^{\infty} A_n\right).$$

Let (B_n) be a decreasing sequence of sets:
Decreasing Sequence:

$$\lim_{n\to\infty} P(B_n) = P\Big(\lim_{n\to\infty} \downarrow B_n\Big) \equiv P\left(\bigcap_{n=1}^{\infty} B_n\right).$$

4 Lebesgue Measure

4.1 Motivation (Number Theory Version)

Focus on half closed interval. Random draws $\omega \in [0,1]$: $\mathbb{P}(\omega \in [0, 0.47)) =?$

- Pick a number: $0.xy\dots$
- Either $x \in \{0,1,2,3\}$ or $x = 4 \cap y \in \{0,1,\dots,0.6\}$.

$$\#\text{ possibilities} = (4 \cdot 10) + (1 \cdot 7) = 47 \text{ out of } (10 \cdot 10) \text{ possibilities.}$$

$$\mathbb{P}(w \in [0, 0.47]) = \frac{47}{100} = 0.47 \quad \text{Some shape of Uniform CDF.}$$

4.2 Construction

Definition 4.2.1 (Lebesgue Measure on $\mathbb{R} \cap [0,1]$).
More Generally, for $a, b \in \mathbb{R} \cap [0,1]$, $a < b$, Lebesgue Measure is a measure defined on the power set of $\mathbb{R} \cap [0,1]$ (Warning! this is ideal but is not true. reason to be discussed in the next chapter.) onto $\mathbb{R}$ such that the mapping is non-negative and sigma-additive. It is defined by the following:

$$\mathbb{P}([0,a)) = a, \quad \mathbb{P}([0,b)) = b.$$

$$\mathbb{P}([a,b)) = \mathbb{P}([0,b)) - \mathbb{P}([0,a)) = b - a.$$

Proposition 4.2.2 Additional Features on Lebesgue Measure:

- Unity: $\mathbb{P}([0,1]) = 1$
- Translational Invariant: $\mathbb{P}(x + A) = \mathbb{P}(A), \ \forall x \in \mathbb{R}$

Proposition 4.2.3 :

1. **Open interval:** $\mathbb{P}((a,b)) = \lim_{n\to 0} \mathbb{P}((a + \frac{1}{n}, b)) = \lim_{n \to 0} b - (a + \frac{1}{n}) = b - a.$
2. **Single element:** $\mathbb{P}(\{a\}) = \mathbb{P}([a,b]) - \mathbb{P}((a,b)) = b - a - (b - a) = 0.$
3. **Closed interval:** $\mathbb{P}([a,b]) = \mathbb{P}([a,b)) + \mathbb{P}(\{b\}) = \mathbb{P}([a,b]) = b - a.$

Remark 4.1 *Using Kolmogorov axioms for measures, $A \subset [0,1]$, finite or countable, $\mathbb{P}(A) = 0$. As a result, $\mathbb{P}(\mathbb{Q} \cap [0,1]) = 0$.*

Outer Measure

Motivation and Definition of Outer Measure

Recall the definition of upper and lower Darboux sum in Riemann integral setting:

$$U(f) = \sum_{i=1}^{N} \sup_{x\in[x_i,x_{i+1}]} f(x) \cdot \underbrace{(x_{i+1} - x_i)}_{\text{length}}$$

Definition of Riemann non-integrable involves the upper integral and the lower integral as two limits do not agree, which likely boils down to some partitions $[x_i, x_{i+1}]$ not well-defined. Therefore, to propose a fix, we are motivated to properly define the "length" of any general subset of $\mathbb{R}$.

Definition 4.2.4 (Length). The length $\ell(I)$ of some open interval $I \subset \mathbb{R}$ is a function defined by

$$\ell(I) = \begin{cases} b-a, & I = (b-a), a < b, a, b \in \mathbb{R} \\ 0, & I = \varnothing, \\ \infty & I = (-\infty, a), a \in \mathbb{R} \\ \infty & I = (a, \infty), a \in \mathbb{R} \end{cases}$$

Then suppose $A \subset \mathbb{R}$. The size of A can at most be the **sum of lengths of a sequence of open intervals** I **whose union contains** A. Taking the infimum of such sums over all possible sequences of I, we obtain the outer measure of A, i.e.

Definition 4.2.5 (Outer Measure, $|A|$). For $A \subset \mathbb{R}$,

$$|A| \equiv \inf \left\{ \sum_{k=1}^{\infty} \ell(I_k) \mid I_k \text{ open, } A \subset \bigcup_{k=1}^{\infty} I_k \right\}$$

Proposition 4.2.6 Finite sets have outer measure 0:
Proof: **??**

4.2.1 Properties of Outer Measure

Proposition 4.2.7 Countable subsets of $\mathbb{R}$ have outer measure 0:
Proof: **??**
Proposition 4.2.8 Order preserving of outer measure:

If $A \subset B \subset \mathbb{R}$, then $|A| \leq |B|$ Proof: **??**

Definition 4.2.9 (Translation). For any $A \subset \mathbb{R}, t \in \mathbb{R}$, the translation $t + A$ is defined by

$$t + A = \{t + a \mid a \in A\}$$

Note that the length function should be translation invariant. Therefore, we obtain the proposition that outer measure is translation invariant.

Proposition 4.2.10 Outer measure is translation invariant:

Suppose $t \in \mathbb{R}$ and $A \subset \mathbb{R}$, then $|t + A| = |A|$.

Proposition 4.2.11 Countable Sub-additivity of outer measure:

Suppose $A_1, A_2, \ldots, \subset \mathbb{R}$. Then

$$\left|\bigcup_{k=1}^{\infty} A_k\right| \leq \sum_{k=1}^{\infty} |A_k|$$

Note that this implies finite sub-additivity which could come handy in proof techniques:

$$|A_1 \cup \cdots \cup A_n| \leq |A_1| + \cdots + |A_n|$$

Outer Measure of Closed Bounded Interval

It is apparent for any closed interval $[a, b]$, we can construct a sequence of open cover $(a - \varepsilon, b + \varepsilon)$ and arbitrarily shrink ε. We obtain $|[a, b]| \leq b - a$. However, the other direction requires completeness of $\mathbb{R}$.

Proposition 4.2.12 $|[a, b]| = b - a$**:**

Suppose $a, b \in \mathbb{R}, a < b$. Then $|[a, b]| = b - a$. Proof: **??**

Proposition 4.2.13 Non-trivial intervals are uncountable:

Every **interval** in $\mathbb{R}$ that contains at least two distinct terms is uncountable.

Proposition 4.2.14 Non-additivity:

$\exists A, B \subset \mathbb{R}$ disjoint such that $|A \cup B| \neq |A| + |B|$.

5 Sigma-Algebra

5.1 Motivation

The Banach-Tarski paradox introduces the fundamental problem in measure theory with uncountable latent space Ω. To be more specific, it is our inability to properly define a measure with the aforementioned axioms.

Proposition 5.1.1 Non-existence of extension of length to all subsets of $\mathbb{R}$:

There does not exist a function μ with all the following properties:

a). $\mu : \mathscr{P}(\mathbb{R}) \to [0, \infty]$

b). $\mu(I) = \ell(I) \forall$ open interval I on $\mathbb{R}$

c). Countable additivity: $\mu(\bigcup_k A_k) = \sum_k \mu(A_k)$ for all disjoint $A_k \subset \mathbb{R}$

d). Translation invariant: $\mu(t + A) = \mu(A)\ \forall A \subset \mathbb{R}$ and $t \in \mathbb{R}$.

Proof: A.3
The only condition we can relax is a). Instead of the entire power set, we define the measure to be only on a subset of the power set, defined as "σ-algebra:

5.2 Setup

Theorem 5.2.1 Sigma Algebra
A subset $\mathscr{F} \subseteq \mathscr{P}(X)$ is called a **sigma algebra** if:

1. $\varnothing, X \in \mathscr{F}$,
2. If $A \in \mathscr{F}$, then $\overline{A} \in \mathscr{F}$,
3. If $(A_i)_{i \in \mathbb{N}} \subseteq \mathscr{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

Definition 5.2.2 (Measurable Set). If $A \in \mathscr{F}$, then A is called an **$\mathscr{F}$-measurable set**.

Remark 5.1 *Examples of sigma algebras:*

- ***Trivial sigma algebra:*** $\mathscr{F} = \{\varnothing, X\}$.
- ***Full power set:*** $\mathscr{F} = \mathscr{P}(X)$.

5.3 Properties

Proposition 5.3.1 Intersection of Sigma Algebras:

The countable intersection of sigma algebras is a sigma algebra. If $\mathscr{F}_i$ is a sigma algebra on X for $i \in I$, then:

$$\bigcap_{i \in I} \mathscr{F}_i \text{ is also a sigma algebra.}$$

Definition 5.3.2 (Sigma Algebra Generated by a Set).

For any $\mathcal{M} \subseteq \mathscr{P}(X)$, the smallest sigma algebra containing $\mathcal{M}$ is denoted $\sigma(\mathcal{M})$ and is called the **sigma algebra generated by** $\mathcal{M}$.

1. Collect all large $\mathscr{F}$ as sigma algebras such that $\mathcal{M} \subseteq \mathscr{F}$.

2. Take their intersection:

$$\sigma(\mathcal{M}) = \bigcap_{\mathcal{M} \subseteq \mathscr{F}} \mathscr{F}.$$

Example 5.3.3 Let $X = \{a, b, c, d\}$ and $\mathcal{M} = \{\{a\}, \{b\}\}$. Then:

$$\sigma(\mathcal{M}) = \{\varnothing, X, \{a\}, \{b\}, \{a, b\}, \{c, d\}, \{a, b, c\}, \{a, b, d\}\}.$$

Theorem 5.3.4 closure property of σ-algebras

For a sigma-algebra $\mathscr{F}$, if $A \in \mathscr{F}$, then $\sigma(A) \subseteq \mathscr{F}$.

Proposition 5.3.5 Sigma Algebra on subsets:

If a σ-algebra on a larger space is defined, then any subset of that space has a σ-algebra by intersecting everything with that subset.

5.4 Borel σ-field on $\mathbb{R}$

Definition 5.4.1 (Borel Sigma Algebra).

Let X be a topological space.

The Borel sigma algebra $\mathscr{B}(X)$ is the sigma algebra **generated by all open sets** of X.

Lemma 5.4.2 Compactness:

All compact subset has a finite measure w.r.t. Borel sigma algebra.

Remark 5.2 *The triplet "(Set, σ-algebra, measure)": $(X, \mathscr{F}, \mu)$ is called a* ***measure space****.*

Remark 5.3 *The Borel sigma algebra is particularly useful in analysis and probability theory:*

- *For continuous random variables, the pre-image of an open set under a continuous mapping $f : \mathbb{R} \to \mathbb{R}$ is measurable since it belongs to $\mathscr{B}(\mathbb{R})$.*
- *$\mathscr{B}(\mathbb{R})$ is the natural sigma algebra for defining measures, such as the Lebesgue measure.*

Definition 5.4.3 (Borel σ-field on $\mathbb{R}^d$). The Borel σ-field on $\mathbb{R}^d$, denoted $\mathscr{B}^d$, is the smallest σ-algebra on $\mathbb{R}^d$ containing all Cartesian products of univariate Borel sets:

$$\prod_{i=1}^{d}(a_i, b_i).$$

6 Measurable maps

6.1 Measurable maps

Definition 6.1.1 (Measurable map).
Let $(\Omega_1, \mathscr{F}_1)$, $(\Omega_2, \mathscr{F}_2)$ be measurable spaces.
A map $f : \Omega_1 \to \Omega_2$ is measurable with respect to $(\mathscr{F}_1, \mathscr{F}_2)$ if:

$$f^{-1}(A) \in \mathscr{F}_1,\ \forall A \in \mathscr{F}_2.$$

iff Pre-image of measurable sets are measurable.

Example 6.1.2 Indicator Function
Consider $\chi_A : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$, where

$$\chi_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{else.} \end{cases}$$

To show χ_A is measurable, check all pre-images:

$$\chi_A^{-1}(\varnothing) = \varnothing, \quad \chi_A^{-1}(\mathbb{R}) = \Omega, \quad \chi_A^{-1}(\{1\}) = A, \quad \chi_A^{-1}(\{0\}) = A^c.$$

Since all are in $\mathscr{F}$, χ_A is measurable.

Example 6.1.3 Suppose $f : (\Omega_1, \mathscr{F}_1) \to (\Omega_2, \mathscr{F}_2)$ and $g : (\Omega_2, \mathscr{F}_2) \to (\Omega_3, \mathscr{F}_3)$. If f and g are measurable, then $g \circ f$ is measurable.
For any $A \in \mathscr{F}_3$,

$$(g \circ f)^{-1}(A) = \underbrace{f^{-1}\underbrace{\left(g^{-1}(A)\right)}_{\in \mathscr{F}_2}}_{\in \mathscr{F}_1}.$$

Since $g^{-1}(A) \in \mathscr{F}_2$ and f^{-1} preserves measurability, $(g \circ f)^{-1}(A) \in \mathscr{F}_1$.

Proposition 6.1.4 :
Let $(\Omega, \mathscr{F})$, $(\mathbb{R}, \mathscr{B})$, and $f, g : \Omega \to \mathbb{R}$ be measurable. Then:

1. $f + g, f - g$ are measurable,

2. $|f|$ is measurable.

Proposition 6.1.5 Measurable and Continuity:
Let $f : X \to \mathbb{R}$ be a continuous function. Then f is measurable with respect to the Borel sigma algebra.

6.2 Probability measure to a random variable (Push-forward)

Definition 6.2.1 (Push-forward measure). Push-forward measure P_X is the measure induced on $\mathbb{R}^d$ via the function X, such that:

$$\forall A \in \mathscr{B}^d,\ P_X(A) = P(X^{-1}(A)).$$

For a probability (measure) space $(\Omega, \mathscr{F}, P)$ and a function (random variable)

$$X : (\Omega, \mathscr{F}, P) \to (\mathbb{R}^d, \mathscr{B}^d, \cdot),$$

There are generally two ways to make X measurable map:

1. Fix $\mathscr{B}^d$ and $\mathscr{F}$, ask if X is measurable.
2. Alter $\mathscr{F}$ according to X such that X is measurable.

6.2.1 Method 1: Validation

Ask whether X is measurable with respect to the fixed σ-algebras:

$$\forall A \in \mathscr{B}^d, \quad X^{-1}(A) \in \mathscr{F}.$$

This directly checks the measurability condition. If satisfied, we say then X as a function is measurable.

- *Non-Measurability* shows by the *same set having 2 different sizes (measures)*

Criterion Check (1):

- For $\Omega \xrightarrow{X} \mathbb{R}^d$, $\forall A \in \mathscr{B}^d$, $X^{-1}(A) \in \mathscr{F}$, then X is measurable.
- BUT: Too hard to validate $\forall A \in \mathscr{B}^d$.

6.2.2 Method 2: Generation

Fix $\mathscr{B}^d$, then construct a σ-algebra $\mathscr{F}_X$ (generated by X) that ensures measurability:

$$X^{-1}(\mathscr{B}^d) := \{X^{-1}(A) \mid A \in \mathscr{B}^d\}.$$

Define:

$$\mathscr{F}_X \equiv \sigma(X) \quad \text{(the } \sigma\text{-algebra generated by } X\text{).}$$

This is the smallest σ-algebra that makes X measurable:

- *Actually, $\sigma(X)$ is the smallest σ-algebra for X to be measurable.*
- $\sigma(X)$ contains all pre-images of $\mathscr{B}^d$ under X.

Example 6.2.2
Let $X(\omega) = \mathbb{I}_A(\omega)$ for some $A \subseteq \Omega$, where $\mathbb{I}_A$ is the indicator function of A. Then the σ-algebra generated by X is:

$$\sigma(X) = \{\varnothing, A, A^c, \Omega\}.$$

This σ-algebra usually represents *information sets* in practice.

Theorem 6.2.3 Measurability Validation
Solution: It is sufficient to check the generators B of $\mathscr{B}$ such that $X^{-1}(B) \subset \mathscr{F}$.

Example 6.2.4
Let $\mathcal{L} = \left\{\prod_{i=1}^{d}(-\infty, x_i] \mid x_i \in \mathbb{R}\right\}$ be a generator of $\mathscr{B}^d$. We have $X^{-1}(\mathcal{L}) \subseteq \mathscr{F}$, then we generate the smallest σ-algebra:

$$\sigma(X^{-1}(\mathcal{L}))$$

We need to show that

$$\sigma(X) \subseteq \sigma(X^{-1}(\mathcal{L})) \subseteq \mathscr{F}.$$

which makes X measurable.
example proof: A.4

7 Stochastic Convergence

7.1 Almost Sure Convergence

For a converging sequence of random variables (as functions) $\{X_n\}_{n\in\mathbb{N}}$ and a their limit X, all map from Ω to $\mathbb{R}$, recall the epsilon definition of pointwise convergence,

$$X_n(\omega) \to X(\omega) \quad \text{if} \quad \forall\varepsilon > 0, \exists N_0 \text{ s.t. } |X_n(\omega) - X(\omega)| < \varepsilon,\ \forall n \geq N_0.$$

However, due to random fluctuations, $X_n(\omega) \in \mathbb{R}$ may deviate for some ω. This requires weaker notions of convergence. Instead, we require only **almost all** ω suffices the above condition. To be more mathematically rigorous, the set of ω s.t. the above condition does not hold has a measure 0 w.r.t. the measure defined by the context. This idea turns into ”almost sure convergence”:

Theorem 7.1.1 Almost Sure Convergence

In a sequence space Ω^N, let $(X_n(\omega))_{n=1}^{\infty}$ be a sequence of random variables $\Omega \to \mathbb{R}$. We say:

$$X_n(\omega) \xrightarrow{\text{a.s.}} X(\omega) \quad \text{if} \quad \Pr\Big(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big) = 1.$$

This means the probability of pointwise convergence holds except for a set of ”abnormal” ω, which has probability 0. It is a point-wise convergence for almost all ω.

Note that the limit could either be random or non-random.

In epsilon language:

$$P[\forall\varepsilon > 0, \exists n_0,\ \forall n \geq n_0,\ |X_n - X| < \varepsilon] = 1$$

Or using the definition of what I call ”good sets”:

$$P\left[\bigcap_{\varepsilon>0} B_\varepsilon\right] = 1, \quad B_\varepsilon \equiv \{\omega \in \Omega : \exists n_0, \forall n \geq n_0, |X_n(\omega) - X(\omega)| < \varepsilon\}$$

If given $X_n \xrightarrow{\text{a.s.}} X$, then $P(B_\varepsilon)$ is by definition zero. Recognising that $\cap_\varepsilon B_\varepsilon$ is a decreasing set as smaller choices of ε makes B_ε more restrictive. Therefore, $P\{\cap_\varepsilon B_\varepsilon\} = 1$.

Then the converse becomes one way of showing a.s. convergence in practice. We need to show that $\forall \varepsilon > 0, P(B_\varepsilon) = 1$. Note that each B_ε can also be written as:

$$\begin{aligned} B_\varepsilon &\equiv \{\omega \in \Omega : \exists n_0, \forall n \geq n_0, |X_n(\omega) - X(\omega)| < \varepsilon\} \\ &= \bigcup_{n_0 \in \mathbb{N}} \underbrace{\bigcap_{n \geq n_0} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\}}_{\equiv C_{n_0}} \end{aligned}$$

Let us walk through the above expression carefully. Larger choice of n_9 makes fewer subsets intersecting in the particular C_{n_0}. Fewer subsets' intersection is likely to be less restrictive therefore is larger. We conclude that the sequence of sets C_{n_0} is an increasing sequence. Therefore, the above expression can be written into:

$$\begin{aligned} B_\varepsilon &= \lim_{n_0 \in \mathbb{N}} \uparrow \bigcap_{n \geq n_0} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\} \\ P(B_\varepsilon) &= \lim_{n_0 \to \infty} P\left[\bigcap_{n \geq n_0} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\}\right] \quad \text{By continuity of measure} \end{aligned}$$

Since all probability measures are bounded above by 1, we have the monotone convergence theorem to conclude that the above limit goes to 1.

We can then focus on the limit for all B_ε. To do this, we use $\varepsilon_k = 1/k \in \mathbb{Q}$ to let ε goes to 0, i.e.

$$\bigcap_{\varepsilon > 0} B_\varepsilon = \bigcap_{k > 0} B_{1/k} = \lim_{k \to \infty} \downarrow B_{1/k}$$

Under proper conditions that if all $B_{1/k}$ has probability measure 1, then $P\left[\bigcap_{\varepsilon > 0} B_\varepsilon\right] = 1$.

7.2 Equivalent definitions of almost sure convergence

There are several equivalent conditions for almost sure convergence listed below:

Theorem 7.2.1 Characterization of A.S. Convergence

$X_n(\omega) \xrightarrow{\text{a.s.}} X(\omega)$ iff:

$$\forall \varepsilon > 0, \quad \begin{cases} \lim_{N \to \infty} \Pr\left(\bigcap_{n=N}^{\infty} |X_n(\omega) - X(\omega)| < \varepsilon\right) = 1, \\ \lim_{N \to \infty} \Pr\left(\bigcup_{n=N}^{\infty} |X_n(\omega) - X(\omega)| \geq \varepsilon\right) = 0, \\ \lim_{N \to \infty} \Pr\left(\sup_{n \geq N} |X_n(\omega) - X(\omega)| \geq \varepsilon\right) = 0. \end{cases}$$

7.3 Convergence in Probability

Theorem 7.3.1 Convergence in Probability

$$X_n(\omega) \xrightarrow{\text{p}} X(\omega) \quad (\text{Or } \operatorname*{plim}_{n\to\infty} X_n = X) \quad \text{if} \quad \forall\varepsilon > 0, \forall\omega \in \Omega \quad \lim_{n\to\infty} \Pr\left(|X_n - X| < \varepsilon\right) = 1.$$

But there might not be a single universal $N(\omega)$ such that for all $n > N(\omega)$, $X_n(\omega)$ stays near $X(\omega)$. The probability that they are close is going to 1, but that probability can come from “different” ω at different times.

Remark 7.1 (Additional distinctions) *You can have $\{X_n\}$ that converge in probability to X, but there is no single set of outcomes of probability 1 on which you get pointwise convergence. In other words, the set of ω’s where it fails to converge pointwise might not be measure zero—it might be a “moving target,” so that for each n only a small subset violates the ε-closeness condition, but across infinitely many n, you don’t get pointwise convergence for a fixed set of ω.*

Lemma 7.3.2 Borel-Cantelli Lemma:

For any sequence of sets $\{A_n\}$, if their infinite sum converges, $\sum_{n\in\mathbb{N}} P(A_n) < \infty$, (which also implies that the probability at the tail goes to 0), then the probability that infinitely many of them occur in the tail is 0, i.e.

$$\Pr\left(\bigcap_{n=N_0}^{\infty} \bigcup_{n=N_0}^{\infty} A_n\right) = 0.$$

Remark 7.2 *Borel-Cantelli Lemma is one way to prove A.S. convergence.*

Proposition 7.3.3 :

A.S. Convergence ⇒ Convergence in Probability.

Proof: A.2.3

7.4 Application: Consistency

Example 7.4.1 Application in Estimator Consistency

Let (X_n) denote a dataset from some sampling of the true population, θ being some true parameter in the population, and let $\widehat{\theta}_n(\omega)$ be an estimator of θ. To show consistency:

$$\widehat{\theta}_n(\omega) = f(X_n(\omega)) \quad \text{and} \quad \widehat{\theta}_n(\omega) \xrightarrow{\text{a.s.}} \theta.$$

Definition 7.4.2 (Strong Consistency). An estimator $\widehat{\theta}_n$ is strongly consistent for θ if:

$$\Pr\left(\lim_{n\to\infty} \widehat{\theta}_n = \theta\right) = 1.$$

The data generating process (sampling process) of X_n converges to the true population X almost surely iff its estimator $\widehat{\theta}$ converges to the true parameter θ almost surely.

Theorem 7.4.3 Law of Large Number

Strong Law of Large Number:

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} p \quad \Leftrightarrow \quad f_n(A) \xrightarrow{\text{a.s.}} P(A).$$

Weak Law of Large Number:

$$\frac{S_n}{n} \overset{\text{p}}{\to} p \quad \Leftrightarrow \quad f_n(A) \overset{\text{p}}{\to} P(A)$$

8 Bayes' Rule

8.1 Conditional Probability

Definition 8.1.1 (Conditional Probability).
Let $A, B \subseteq \Omega$. Conditional probability measures how the occurrence of B influences $P(A)$. If $P(B) > 0$

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

Proposition 8.1.2 Partition of Ω:
Assume $\{H_i\}_{i=1}^{\infty}$ is a partition of Ω (disjoint subsets such that $\bigcup_{i=1}^{\infty} H_i = \Omega$). Then for any $B \subseteq \Omega$:

$$P(B) = \sum_{i=1}^{\infty} P(B \cap H_i) = \sum_{i=1}^{\infty} P(B \mid H_i)P(H_i).$$

8.2 Bayes' Rule and Problem

Bayes' rule is used to update probabilities of causes given data.

- $P(H_i)$: Prior probability of potential cause $H_i, i = 1, 2, \ldots$.
- $P(A \mid H_i)$: Probability of event A if H_i is the cause.
- $P(H_i \mid A)$: Posterior probability of H_i given A.

Using Bayes' Rule:

$$P(H_i \mid A) = \frac{P(A \mid H_i)P(H_i)}{\sum_{j=1}^{\infty} P(A \mid H_j)P(H_j)} \propto P(A \mid H_i)P(H_i).$$

8.3 Independence

Let $A, B \subseteq \Omega$. Events A and B are independent if:

$$P(A \mid B) = P(A) \quad \text{or equivalently} \quad P(A \cap B) = P(A)P(B).$$

- Independence holds only if $P(B) > 0$.
- If $P(B) = 0$, the independence condition $P(A \cap B) = P(A)P(B)$ is not well-defined.

Remark 8.1 *If A, B are incompatible ($A \cap B = \varnothing$), they cannot be independent unless $P(A) = 0$ or $P(B) = 0$.*

8.4 Chain Rule of Independence

The chain rule for independence states:

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

This is particularly useful for time series analysis and Markov processes.

8.5 Mutual and Conditional Independence

Definition 8.5.1 (Mutual Independence). The events $A_1, A_2, \ldots, A_n$ are mutually independent if:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Mutual independence does not imply pairwise independence.

Definition 8.5.2 (Conditional Independence). Let $A, B, C \subseteq \Omega$. Events A and B are conditionally independent given C if:

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

Conditional independence does not imply pairwise independence.

Definition 8.5.3 (Mutual Independence of a Family of Events). Let $I \subset \mathbb{N}$ index a family of events $\{A_i\}_{i\in I}$. The events are mutually independent if:

$$P\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} P(A_i), \quad \forall J \subseteq I.$$

9 Combinatorics and Estimating Proportions

We consider a non-parametric estimator $\widehat{\theta}$ for proportions:

- Sampling can be done **with replacement** or **without replacement**.
- Example: Bootstrap sampling is an example of sampling with replacement.

To compute the probabilities, let N be the total population size and n the sample size. Uniform sampling leads to two cases:

- **With replacement:** Uniform distribution on $\Omega = \{1, 2, \ldots, N\}^n$.
- **Without replacement:** Uniform distribution on $\Omega = \binom{N}{n}$.

Remark 9.1 *If n is too small for N, then $\binom{N}{n} \neq N^n$, as ordering plays a role.*

Ordered vs. Unordered Samples

- **Ordered:** The total number of possible ordered samples without replacement is:

$$P(N, n) = \frac{N!}{(N-n)!}.$$

- **Unordered:** For unordered sampling without replacement, the total number is:

$$\binom{N}{n} = \frac{P(N, n)}{n!}.$$

Uniformity Assumption When sampling is uniform:

- **Without replacement:** All samples of size n have the same probability.
- **With replacement:** Uniformity may not hold; for example, when $N = 2$ and $n = 2$, ordered samples (a, b), (b, a), etc., do not all have the same probability.

Specific Types in Ordered Samples Suppose we have R red balls in a total of N objects. We want to calculate the number of samples that include k red balls:

- Choose the k locations for the red balls:

$$\binom{n}{k}.$$

- Choose the remaining $n-k$ white balls:

$$\binom{N-k}{n-k}.$$

Results

- **With replacement:**

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \sim \text{Binomial}(n, p).$$

- **Without replacement:**

$$P(S_n = k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}, \quad \sim \text{Hypergeometric}(N, n, R).$$

Remark 9.2 *Both results define probability mass functions (p.m.f.).*

Maximum Likelihood Estimation (MLE) Definition 9.0.1 (MLE). Given that we observe k successes in n trials, we estimate p by:

$$\widehat{p} = \arg\max_p P(S_n = k).$$

Bayesian Estimation Bayesian estimation adds a prior distribution for the parameter p:

- **Uniform Prior:** Assume a prior distribution for p is uniform on $[0, 1]$, leading to:

$$P(S_n = k) = \frac{\binom{n}{k} \int_0^1 p^k (1-p)^{n-k} \, dp}{\int_0^1 \, dp}.$$

- Use Bayes' rule to calculate posterior probabilities:

$$P(p \mid S_n = k) \propto P(S_n = k \mid p) P(p).$$

Predictive Probability

Theorem 9.0.2 Laplace's law of succession

Laplace's law of succession predicts the probability of a future success given past data:

$$P(S_{n+1} = k+1 \mid S_n = k) = \frac{k+1}{n+2}.$$

Remark 9.3 *This result assumes we "pretend" to have observed one success and one failure before starting the experiment.*

10 Cumulative Density Function

10.1 Cumulative Distribution Function

For $X : (\Omega, \mathscr{F}, P) \to \mathbb{R}^d$ as a random variable, we define the **probability distribution** P_X as the push-forward:

$$\forall A \in \mathscr{B}^d, \quad P_X(A) = P(X^{-1}(A)) = P(X \in A).$$

From Theorem 8.1, we just need to check the generators:

$$A \in L^d \equiv \left\{ \prod_{i=1}^{d} (-\infty, x_i],\ x_i \in \mathbb{R} \right\}.$$

Then it's natural to describe P_X through the CDF F_X. Define:

$$\forall x \in \mathbb{R}^d, \quad F_X(x) = P_X\left(\prod_{i=1}^{d} (-\infty, x_i] \right) = P(X_1 \le x_1, \dots, X_d \le x_d).$$

Example 10.1.1

1. $P_X((-\infty, x_1]) = P_X(X \le x_1) = F_X(x_1)$.
2. $P_X((-\infty, x)) = P(X < x)$:

$$= P\left(\bigcup_{n=1}^{\infty} (-\infty, x - \tfrac{1}{n}] \right) = \lim_{n\to\infty} P((-\infty, x - \tfrac{1}{n}) = F_X(x).$$

3. $P_X(\{x\}) = P(X = x) = F_X(x) - F_X(x^-)$.

Since countable limits only, the CDF can have only countably many jumps: $I := \{x \in \mathbb{R} : F_X(x) > F_X(x^-)\},\ |I|$ countable.

Three Simple Cases:

1. Having only jumps:

$$P(X \in \mathbb{Z}) = \sum_{x \in \mathbb{Z}} [F_X(x) - F_X(x^-)] = 1.$$

   Then X is a **discrete random variable**.

2. $P_X(X \le 1) = \sum_{x \le 1} [F_X(x) - F_X(x^-)] \le 1$:

$$P_X(A) = P_X(A \cap I) \le 1 \quad \forall A \text{ countable}.$$

3. **Extreme Case:** The following are equivalent:
   (a) $F_X : \mathbb{R} \to [0,1]$ is continuous.
   (b) $P_X(X = x) = 0 \quad \forall x \in \mathbb{R}$.
   (c) X is a **continuous random variable**.

Proposition 10.1.2 CDF:

1. **Cadlag**: Right-continuous.
2. Non-decreasing.
3. Sums up to 1: $F_X(-\infty) = 0,\ F_X(\infty) = 1$.

Theorem 10.1.3 Theorem of Push-Forward Distribution
If X, Y are random variables $X, Y : \Omega \to \mathbb{R}^d$, then the push-forward $P_X = P_Y$ if and only if $F_X = F_Y$.

Remark 10.1 *For $\mathscr{B}$ on $\mathbb{R}$ because of Section 9, the same generators apply.*
(CDF is meaningful on the generators.)

10.2 Higher Dimensions:

Theorem 10.2.1 Theorem 3
Recall the properties of CDF in 1 dimension: There exists a real random variable $X : \Omega \to \mathbb{R}$ such that F is the CDF of X *if and only if* F fulfills the 3 properties of a CDF:

$$P((-\infty, x]) = F_X(x) \quad \forall x.$$

Remark 10.2 *Higher Dimension*
The above theorem is necessary yet not sufficient for a multidimensional CDF. The following conditions are needed.

In $\mathbb{R}^2$, $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$, i.e., a *joint CDF.*

To be a CDF of some joint (X, Y), the following must be true:

1. $F_{X,Y}(x, y) : \mathbb{R}^2 \to [0,1]$ is finite.
2. **Right continuity:** $F_{X,Y}(x^-, y),\ F_{X,Y}(x, y^-)$ are non-decreasing $\forall y \in \mathbb{R}$.
3. $\lim_{x\to\infty} F(x, y^n) = F(x, y) \quad \forall y_n \to y$.

4. $F_{X,Y}(-\infty, \infty) = 0,\ F_{X,Y}(\infty, \infty) = 1$.

5. **Additivity:** $\forall x, y \in \mathbb{R},\ h \geq 0$:

$$F_{X,Y}(x+h, y+k) - F_{X,Y}(x+h, y) - F_{X,Y}(x, y+k) + F_{X,Y}(x, y) \geq 0.$$

10.3 Marginal CDFs:

If $F_{X,Y}(x, y)$ is the CDF for (X, Y), then define:

$$F_X(x) = F_{X,Y}(x, \infty), \quad F_Y(y) = F_{X,Y}(\infty, y).$$

(Projection, idempotent.)

Independent Random Variables:

For $i = 1, \ldots, n$, $X_i : \Omega \to \mathbb{R}^{d_i}$ with:

$$\sigma(X_i) = X^{-1}(\mathscr{B}^{d_i}) = \{X_i^{-1}(B_i) : B_i \in \mathscr{B}^{d_i}\} \text{ on } \Omega.$$

Theorem 10.3.1 TFAE (The Following Are Equivalent):

1. $\forall A_i \in \sigma(X_i),\ \{A_i\}_{i=1}^n$ are mutually independent.
2. $\forall B_i \in \mathscr{B}^{d_i},\ P\left(\omega : X_i(\omega) \in B_i, i = 1, \ldots, n\right) = \prod_{i=1}^n P(X_i \in B_i)$.
3. $\forall x_i \in \mathbb{R}^{d_i},\ F_{X_1,\ldots,X_n} = \prod_{i=1}^n F_{X_i}$.
4. $\forall B_i \in \mathscr{B}^{d_i},\ P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^n P(X_i \in B_i)$.

11 Probability Density Function

11.1 Univariate Density

Recall the cumulative distribution function (CDF) is defined as:

$$P((-\infty, x]) = F_X(x).$$

- If X is continuous, then $F_X(x)$ is continuous.
- *Is there a* $f_X(x)$ *such that* $\forall x \in \mathbb{R}, F_X(x) = \int_{-\infty}^{x} f_X(t)dt$, *being a measure?*

  Then $f_X(x)$ is called the **PDF**.

Sufficient Condition: If F_X is continuously differentiable $\forall x \in \mathbb{R}$:

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

Definition 11.1.1 (Absolute continuity).
F_X is absolutely continuous if and only if there exists $f_X(x)$ such that:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \quad \text{(Lebesgue integral to be introduced later).}$$

Remark 11.1 *Following remarks*

R1. Differentiability $\Rightarrow$ *Absolute continuity* $\Rightarrow$ *Continuity.*

R2. Properties of $f_X(x)$*:*

1. $f_X(x) \geq 0, \ \forall x \in \mathbb{R}$,
2. $\int_{-\infty}^{\infty} f_X(t)\,dt = 1$.

R3. $F_X(x)$ *is continuously differentiable if and only if points where* $f_X(x)$ *is continuous* ***(by Fundamental Theorem of Calculus)****.*

R4. When it exists, $f_X(x)$ *is unique almost everywhere with respect to Lebesgue measure and can be computed by* $\frac{\mathrm{d}}{\mathrm{d}x}F_X(x)$*. ideas from functional analysis weakens this condition.*

Example 11.1.2 If $X \sim \text{Uniform}(a, b)$ with Lebesgue measure:

$$f_X(x) = \mathbb{I}_{(a,b)}(x) \cdot \frac{1}{b-a}.$$

- Support: $\text{supp}(f_X(x))$ is where $f_X(x) > 0$.
- The Lebesgue measure gives the shape of the measure.

For small $h > 0$, we approximate:

$$P(X \in [x, x+h]) \approx f_X(x) \cdot h \quad \Rightarrow \quad \text{Riemann Integral:}$$

11.2 Multivariate Probability Density Function

For a joint density $f_X(x) : \mathbb{R}^d \to \mathbb{R}$, we define:

$$F_X(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(t_1, \ldots, t_d) \, dt_1 \cdots dt_d.$$

Here, $f_X(x)$ is the joint density, and F_X is the joint CDF.

Remark 11.2 *Multivariate PDF*

- *Up to Lebesgue measure zero, the density $f_X(x)$ satisfies:*

$$\frac{\partial^d F_X}{\partial x_1 \cdots \partial x_d} = f_X(x_1, \ldots, x_d).$$

- *For small changes $h > 0$, we approximate:*

$$P(X \in [x, x+h], Y \in [y, y+h]) \approx f_{X,Y}(x, y) \cdot h \cdot k.$$

Theorem 11.2.1 Marginal Density

If (X, Y) is absolutely continuous, then:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \quad \forall x \in \mathbb{R}.$$

(Symmetry works.)

11.3 Independence of PDFs

Theorem 11.3.1 Independence

- If X and Y are absolutely continuous and $X \perp Y$, then the joint density $f_{X,Y}$ is multiplicatively separable:

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

- If (X,Y) is absolutely continuous with joint density $f_{X,Y}(x,y) = g(x) \cdot h(y)$, then $X \perp Y$, and:

$$f_X(x) = c \cdot g(x), \quad f_Y(y) = c \cdot h(y) \quad \text{for some constant } c > 0.$$

12 Transformation of Density

12.1 Univariate:

- $X : (\Omega, \mathscr{B}, P) \to \mathbb{R}$ with density $f_X : \mathbb{R} \to \mathbb{R}^+$ such that $f_X(x) = \mathbb{I}_{(a,b)}(x) \cdot f(x)$.
- Assume φ is injective and continuous: $(a, b) \to (c, d)$. *Need the same dimension.*
- **Question:** Define $Y = \varphi(X)$:
  1. Is Y absolutely continuous?
  2. If yes, what is f_Y?

Theorem 12.1.1 Transformation

- If f_X has a density and φ is invertible (for now),
- Sufficient condition: φ is strictly monotone and strictly continuous.
- $\Rightarrow$ From now, we deal with *strictly monotone functions.*

1) φ strictly increasing:

$$\text{Suppose } \varphi : (a, b) \to \mathbb{R},\ \varphi(a) = c,\ \varphi(b) = d, \text{ then:}$$

1. $F_Y(y) = 0 \quad \forall y \leq c$ and $F_Y(y) = 1 \quad \forall y \geq d$,
2. $\forall y \in (c, d), F_Y(y) = P(X \leq \varphi^{-1}(y)) = F_X(\varphi^{-1}(y))$,
3. $\forall y \in (c, d),$ if $F_X, \varphi \in \mathscr{C}^1,$ then $F_Y \in \mathscr{C}^1$

The above three conditions guarantee a well-defined f_Y.

Remark 12.1

- *General φ can be partitioned into countable intervals in which φ is strictly increasing/decreasing $\Rightarrow$ bounded variation.*
- *Exception: Brownian motion, Weierstrass function*

12.2 Density of Y:

$$\text{Then: } f_Y \equiv \frac{dF_Y}{dy} = \mathbb{I}_{\{\varphi(X) \leq y\}}(y) \cdot f_X(\varphi^{-1}(y)) \cdot \left| \frac{d\varphi^{-1}(y)}{dy} \right|^{-1}.$$

Lemma 12.2.1 Special Case of Push-Forward::
If $\exists f_X, \varphi^{-1}$, φ^{-1} differentiable, and if F_X absolutely continuous, then F_Y is absolutely continuous

Jacobian Formula:

Theorem 12.2.2 Jacobian Formula
If $X : (\Omega, \mathscr{F}, P) \to \mathbb{R}^d$ works for f_X, s.t.:

$$\forall x \in \mathbb{R},\ f_X(x) = \mathbb{I}_{(a,b)}(x) \cdot f(x) \text{ and } \varphi \in \mathscr{C}^1(U, V) \text{ diffeomorphism, } U, V \in \mathbb{R}^d$$

then for $d = 1$, $y = \varphi(X)$ is absolutely continuous with density:

$$f_Y(y) = \mathbb{I}_{(c,d)}(y) \cdot f_X(\varphi^{-1}(y)) \cdot \left| \frac{d}{dy} \varphi^{-1}(y) \right|^{-1}.$$

And for $d > 1$, $J_{\varphi^{-1}}$ nonsingular, then we have $y = \varphi(X)$ absolutely continuous with density:

$$f_Y(y) = \mathbb{I}_V(y) f_X(\varphi^{-1}(y)) \cdot \left|\det J_{\varphi^{-1}}(y)\right|^{-1}.$$

12.3 Convolution:

- If $X, Y : (\Omega, \mathscr{F}, P) \to \mathbb{R}^d$ are independent and absolutely continuous,

$$Z \equiv X + Y \text{ is absolutely continuous.}$$

The density f_Z is called the **convolution** of X and Y:

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

Trick: Find (X, Z) such that $(X, Z) \xrightarrow{\varphi} (X, Y)$, then φ is invertible. Then:

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

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

13 Lebesgue Integration

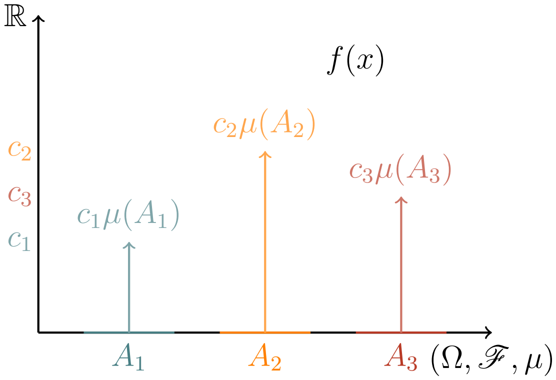
13.1 Set up:

- In $(\Omega, \mathscr{F}, \mu)$, let $f : \Omega \to \mathbb{R}$ be measurable.
- **Recall:** Indicator functions are measurable.
  - Integration of indicator functions equals the measure of the indicator function.

Definition 13.1.1 (Simple Functions).
f is said to be a simple function is f is a finite linear combinations of indicator functions:

$$f(x) = \sum_{i=1}^{n} c_i \cdot \mathbb{I}_{A_i}(x), \quad \text{for some measurable sets } \{A_i\} \text{ and constants } c_i \in \mathbb{R}.$$

Example 13.1.2 Simple functions



Definition 13.1.3 (Lebesgue Integral (for Simple Functions)).
Define the set

$$S^+ := \{f : \Omega \to \mathbb{R} \mid f \text{ is simple function, } f \geq 0\}.$$

For $f \in S^+$: define its Lebesgue integral w.r.t. its measure μ as:

$$\int_\Omega f \, d\mu := \sum_{i=1}^{n} c_i \cdot \mu(A_i), \quad \forall f \in S^+$$

Notation:

$$\int_\Omega f \, d\mu \quad \text{or} \quad \int_\Omega f(x) \, d\mu(x).$$

Proposition 13.1.4 Lebesgue Integral:

- **Linearity:** $\int_\Omega (\alpha f + \beta g)\, d\mu = \alpha \int_\Omega f\, d\mu + \beta \int_\Omega g\, d\mu, \quad \forall f, g \in S^+, \alpha, \beta \geq 0.$
- **Monotonicity:** If $f \leq g$, then $\int_\Omega f\, d\mu \leq \int_\Omega g\, d\mu$.

13.2 Lebesgue Integral

Definition 13.2.1 (Lebesgue Integral).

1. For any $f : \Omega \to \mathbb{R}$, define the set $\mathcal{S_\{}^+ = \{h \in S^+ \mid h < f\}$.
2. Compute all Lebesgue integral $I(h)$ according to the previous definition.

We define

$$\int_\Omega f d\mu = \sup_{h \in \mathcal{S_\{}^+} I(h)$$

And we note that f is μ-integrable (or just integrable if the context if clear) if $\int_\Omega f d\mu < \infty$.

13.3 Important results and theorems

Lemma 13.3.1 Fatou's Lemma:

Given a measure space $(\Omega, \mathscr{F}, \mu)$ and a sequence of measurable, non-negative functions $\{f_n\}$ each maps from $(\Omega, \mathscr{F}, \mu) \to (\mathbb{R}, \mathscr{B}, \cdot)$. Define a function $f \equiv \liminf_{n\to\infty} f_n(x)$, then f is measurable and we have the following inequality:

$$\int_\Omega f\, d\mu \leq \liminf_{n\to\infty} \int_\Omega f_n\, d\mu$$

Corollary 13.3.2 Beppo Levi's Monotone Convergence:

Let $(\Omega, \mathscr{F}, \mu)$ be a measure space, and let $\{X_n\}$ be a sequence of non-negative measurable functions defined on Ω. Suppose:

$$0 \leq X_n(\omega) \leq X_{n+1}(\omega) \quad \forall \omega \in \Omega \quad \text{(monotonically increasing sequence of functions)}$$

Define the pointwise limit function:

$$X(\omega) = \lim_{n\to\infty} X_n(\omega), \quad \forall \omega \in \Omega$$

Then:

$$\lim_{n\to\infty} \int_\Omega X_n\, d\mu = \int_\Omega X\, d\mu$$

Theorem 13.3.3 Lebesgue's Dominated Convergence Theorem

Let $(\Omega, \mathscr{F}, \mu)$ be a measure space, and let $\{f_n\}$ be a sequence of measurable functions mapping from $(\Omega, \mathscr{F}, \mu)$ to $(\mathbb{R}, \mathscr{B})$. Suppose:

- $f_n \to f$ pointwise almost everywhere on Ω, and
- there exists an integrable function $g\colon \Omega \to \mathbb{R}$ such that $|f_n(\omega)| \leq g(\omega)$ for all $\omega \in \Omega$ and all $n \in \mathbb{N}$.

Then all f_n and f integrable, and

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega \lim_{n\to\infty} f_n d\mu = \int_\Omega f d\mu$$

13.4 Expectation

Definition 13.4.1 (Expectation). Consider a measure space $(\Omega, \mathscr{F}, P)$ and a random variable $X : \Omega \to \mathbb{R}$. Its expectation is defined to be a Lebesgue integral

$$\mathbf{E}[X] = \int_\Omega X(\omega) dP(\omega)$$

Corollary 13.4.2 Jensen's Inequality:

If X is a random variable and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, then:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Special Case: For $\varphi(x) = |x|$, we obtain:

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$$

Intuition: The expectation of a convex function applied to X is at least as large as applying the function to the expectation of X . Convexity "pulls the curve upwards," leading to this inequality.

Corollary 13.4.3 Markov's Inequality:

For a non-negative random variable X and any $\alpha > 0$.

$$P(|X| > \alpha) \leq \frac{\mathbb{E}[|X|]}{\alpha}$$

Intuition: The probability that X exceeds some threshold α is bounded by the ratio of its expected value to α. It provides an upper bound on tail probabilities.

14 Moment Generating Functions and Characteristic Functions

14.1 Moment Generating Functions

Definition 14.1.1 (MGF).
Given $X : \Omega \to \mathbb{I} \subset \mathbb{R}$ with a measure space $(\Omega, \mathscr{F}, P)$ where $|\mathbb{I}|$ is countable, the **Moment Generating Function (MGF)** $m_X(t)$ is defined as:

$$m_X(t) \equiv \mathbb{E}[e^{tX}]$$

Remark 14.1 *It is possible for $m_X(t) = \infty$. In fact, very few random variables have MGF.*

Remark 14.2 *If:*

1. *There exists $t \neq 0$ such that $m_X(t)$ is finite for $t \in (-\varepsilon, \varepsilon)$ for some small ε.*

2. *$m_X(t)$ is differentiable k times at $t = 0$.*

Then:

$$\mathbb{E}[X^k] = \left.\frac{\partial^k m_X(t)}{\partial t^k}\right|_{t=0}.$$

14.2 Characteristic Function:

Definition 14.2.1 (Characteristic Function).
The **Characteristic Function (CF)** $\varphi_X : \mathbb{R} \to \mathbb{C}$, of a random variable X is defined as:

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

It is the complex analogue of the MGF

Proposition 14.2.2 Properties of CF:

- $\exp(itX)$ is bounded by the unit circle since $|e^{itX}| = \cos(tX) + i\sin(tX) \leq 1$

- The characteristic function $\varphi_X(t)$ is always well-defined.

- Note that if X is multidimensional, tX is defined by the inner product.

Theorem 14.2.3
If $\mathbb{E}[|X|^k] < \infty$, then $\exists \varphi^{(k)}(t)$ for $k = 1, \ldots, n$ $\forall t \in \mathbb{R}$, and

$$\left.\frac{\partial^k \varphi_X(t)}{\partial t^k}\right|_{t=0} = i^k \mathbb{E}[X^k].$$

This is true by Lebesgue's DCT, given the characteristic function is always bounded by the unit circle therefore the limit (from differentiation) and the integral sign are always interchangeable.

14.3 Convergence in Distribution:

Definition 14.3.1 (Convergence in Distribution).
A sequence of random variables X_n converges in distribution to X (denoted $X_n \xrightarrow{d} X$) if:

$$\forall t \in \mathbb{R}, \quad \varphi_{X_n}(t) \to \varphi_X(t).$$

Where φ is the characteristic function of the push-forward measure.

Remark 14.3 *Note that convergence in distribution happens in the codomain, whereas the aforementioned almost surely convergence and convergence in probability happen in the domain (latent space Ω).*

Proposition 14.3.2 TFAE:

1. $\forall t \in \mathbb{R},\ \varphi_{X_n}(t) \to \varphi_X(t)$.
2. $\forall \varphi \in \mathscr{C}_b(\mathbb{R}),\ \mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$.
3. The cumulative distribution functions converge at all points of continuity:

$$F_n(x) \to F(x).$$

Theorem 14.3.3 Levy's Theorem
A sequence X_j of n-variate random variables converges in distribution to random variable X if and only if the sequence φ_{X_j} converges pointwise to a function φ which is continuous at the origin. Where φ is the characteristic function of the limit X.

The definition $\Leftrightarrow$ the first equivalent condition.

Note: Convergence in distribution $X_n \xrightarrow{d} X$ is weaker than convergence in probability or almost sure convergence:

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X.$$

14.4 Characterization of a Distribution:

Theorem 14.4.1 Characterization of a Distribution

Two distributions P_X and P_Y are equal if and only if their characteristic functions are equal:

$$P_X = P_Y \iff \varphi_X(t) = \varphi_Y(t) \quad \forall t \in \mathbb{R}.$$

* Knowing the characteristic function knows all the moments, knowing all the moments knows the distribution.

14.5 Application: De-Convolution

If X_1 and X_2 are independent random variables:

$$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t) \cdot \varphi_{X_2}(t).$$

Given $X_n = X + V$, if we know φ_{X_n} and φ_V, we can recover the distribution of X:

$$P_X = \frac{\varphi_{X_n}}{\varphi_V}.$$

15 The General Prediction Problem

15.1 Hilbert Space

Note that Hilbert space and functional analysis is not the focus of this note. Just a brief introduction here.

Definition 15.1.1 (Hilbert Space).

A **Hilbert space** $\mathscr{H}$ is a vector space equipped with an inner product $\langle \cdot, \cdot \rangle$, which induces a norm $\|x\| = \sqrt{\langle x, x \rangle}$, such that $\mathscr{H}$ is:

- **Complete:** Every Cauchy sequence in $\mathscr{H}$ converges to a limit within $\mathscr{H}$.
- **Linear:** Closed under vector addition and scalar multiplication. Note that the **function space** is linear, not the functions themselves necessarily.
- **Inner Product Space:** Equipped with an inner product $\langle \cdot, \cdot \rangle$ satisfying symmetry, linearity, and positivity.

Note that if we define an inner product $\langle f, g \rangle$ in squared integrable functions as

$$\langle f, g \rangle = \int_{\Omega} f \cdot g \, d\mu$$

Then the space of squared integrable functions, denoted by $L^2(\mu)$, is a Hilbert space.

15.2 General Prediction Problem

- Goal: Predict a real-valued random variable Y based on random covariates $X \in \mathbb{R}^d$ using a function $h(X)$.
- Assumptions: The prediction function h belongs to a linear vector space $\mathcal{V}$.
- Maintained Assumptions: $h, Y \in \mathcal{V}$ if $\lambda g + \mu h \in \mathcal{V}$ and $1 \in \mathcal{V}$.

We have the following optimisation problem w.r.t. the **mean square error**:

$$\min_{h \in \mathcal{V}} \mathbb{E}\big[(Y - h(X))^2\big].$$

The choice of mean square error as the objective function for this minimization problem is a convenient discretion since it induces a Hilbert space.

Theorem 15.2.1 Conditional Expectation
The function $h^*(X)$ that minimizes the MSE satisfies:

$$h^*(X) = \arg\min_{h \in \mathscr{H}} \mathbb{E}\left[(Y - h(X))^2\right]$$

Is the conditional expectation function (to be defined later)

$$h^*(X) = \mathbf{E}[Y|X]$$

Characterized by

$$Y = \mathbf{E}[Y|X] + \varepsilon$$

With

- $\mathbb{E}[\varepsilon] = 0$ (zero mean),
- $\mathrm{Cov}(\varepsilon, h(X)) = 0$, $\forall h \in \mathscr{H}$.

Remark 15.1 *MSE can be decomposed as variance + bias:*

$$\mathbb{E}\left[(Y - h(X))^2\right] = \underbrace{Var[Y - h(X)]}_{Variance} + \Big(\underbrace{\mathbb{E}[Y] - \mathbb{E}[h(X)]}_{Bias}\Big)^2.$$

Since $h \in \mathscr{H}$ contains the unit element therefore all constant functions, we can set the bias equal to zero without affecting the variance to a certain extent.

15.3 Optimality Conditions

- **Necessary condition for optimality:** $h^*(X)$ is optimal if:

$$\mathrm{Cov}(Y - h^*(X), h(X)) = 0 \quad \forall h \in \mathcal{V}.$$

- **Sufficient condition for optimality:**

$$\mathbb{E}\left[(Y - h^*(X))h(X)\right] = 0.$$

Theorem 15.3.1

- *If $h^*(X) \in \mathscr{H}$, it is unique almost surely.*
- *If $\mathscr{H}$ is a closed subspace of $L^2(P)$, then $h^*(X)$ exists and is optimal.*

Example 15.3.2 Affine Regression
Let $\mathcal{V}$ be the vector space spanned by $1, X_1, \dots, X_d$. Then:

$$h^*(X) = a + b^\top X, \quad \text{where} \quad \begin{cases} a = \mathbb{E}[Y] - b^\top \mathbb{E}[X], \\ b = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y). \end{cases}$$

Solution:

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y.$$

And

$$X\widehat{\beta} = \underbrace{X(X^\top X)^{-1}X^\top}_{\equiv P \text{ Projection matrix}} Y.$$

Projection Interpretation

$$h^*(X) = \mathbb{E}[Y|X]$$

is the *unique argument* minimizing the MSE. The solution can be interpreted as a projection of Y onto $\mathscr{H}$, a subspace of $L^2(\mathbb{P})$, which is a Hilbert space.

15.4 Properties

There is **no** full closed form for $\mathbb{E}[Y \mid X]$. All we know is:

- $\mathbb{E}[Y \mid X]$ is the only function in $L^2(P)$ of X such that:

$$\mathbb{E}[Yh(X)] = \mathbb{E}[\mathbb{E}[Y \mid X]h(X)] \quad \forall h(X) \in L^2(P).$$

- $\mathrm{cov}(Y, h(X)) \iff \mathrm{cov}(Y, h(X)) = 0$, so anything outside the $L^2(P)$ differs by an orthogonal function $\implies$ doesn't matter.

15.4.1 Discrete Case

$$\mathbb{E}[Y \mid X = x] : \mathbb{R} \to \mathbb{R}, \text{ thus depends on } x.$$

Claim 15.4.1:

$$\mathbb{E}[Y \mid X = x] = \sum_y y \cdot P_{Y|X}(Y = y|X = x) \equiv \psi(x)$$

where

$$P_{Y|X}(Y \mid X = x) = \frac{P(\{Y = y\} \cap \{X = x\})}{P(X = x)}.$$

16 The Gaussian Distribution

17 Introduction to Asymptotic Theory

A Some Proofs

A.1 Discrete Probability Measure

A.1.1 Boole's Inequality

Statement: $P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n)$

Proof. We first construct a sequence of disjoint sets B_n that each set in the sequence is defined the following:

$$B_1 = A_1, B_2 = A_2 \backslash A_1, B_3 = A_3 \backslash (A_2 \cup A_1), \dots B_n = A_n \backslash \left(\bigcup_{i=1}^{n-1} A_i \right)$$

Claim A.1.1:

All B_n disjoint.

To show this, fix $m, k \in \mathbb{N}$ and observe B_m and B_k. Without loss of generality, suppose that $m < k$. Fix $b \in B_m$, then $b \in A_m$. We can also write $B_k = A_k \backslash \left(\bigcup_{i=1}^{k-1} A_i \right) = A_k \backslash \left(\bigcup_{i=1}^{m-1} A_i \cup A_m \cup \bigcup_{i=m+1}^{k-1} A_i \right)$ which has A_m cut out in the second part. Therefore, $b \notin B_k$.

Similarly, fix $c \in B_k$, then $c \notin A_m$, but $\forall b \in B_m, b \in A_m$. So $c \notin B_m$. We have shown that all B_n disjoint.

Therefore, for any two B_m, B_k, $B_m \cap B_k = \varnothing$. We have all B_i disjoint.

Claim A.1.2:

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

To show this, $\bigcup_{n=1}^{\infty} B_n \subseteq \bigcup_{n=1}^{\infty} A_n$ is trivial as all $B_n \subset A_n$ by construction. We now need to show the converse is true as well. Fix $a \in \bigcup_{n=1}^{\infty} A_n$, then $\exists$ some m such that $a \in A_m$, then we have two scenario:

Scenario 1: $a \in A_m \backslash \left(\bigcup_{i=1}^{m-1} A_i \right)$

Then $a \in B_m$ which implies $a \in \bigcup_{n=1}^{\infty} B_n$, we have $\bigcup_{n=1}^{\infty} A_n \subseteq \bigcup_{n=1}^{\infty} B_n$.

Scenario 2: $a \notin A_m \backslash \left(\bigcup_{i=1}^{m-1} A_i \right)$.

This implies that $a \in \left(\bigcup_{i=1}^{m-1} A_i \right)$, then $\exists k < m$ such that $a \in A_k$. We could again break it into the above two scenarios by replacing m with the new k. This process will not be infinite as falling into scenario 2 will give us a new index that is strictly smaller than the previous one. Since $n \in \mathbb{N}^+$, a has to belong to some $A_l \backslash \left(\bigcup_{i=1}^{l-1} A_i \right) = B_l$. Then $a \in \bigcup_{n=1}^{\infty} B_n$. We have shown that $\bigcup_{n=1}^{\infty} A_n \subseteq \bigcup_{n=1}^{\infty} B_n$, therefore $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$.

Since $B_n \subseteq A_n \; \forall n$, we have the monotonicity of the measures:

$$\begin{gathered} P(B_n) \leq P(A_n) \\ \sum_{n=1}^{\infty} P(B_n) \leq \sum_{n=1}^{\infty} P(A_n) \end{gathered}$$

By Kolmogorov axiom 3, we have $P(\bigcup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n) \; \forall B_n$ disjoint, we have:

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

By above claim 2, we have $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$, then

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n) &\leq \sum_{n=1}^{\infty} P(A_n) \\ P\left(\bigcup_{n=1}^{\infty} A_n\right) &\leq \sum_{n=1}^{\infty} P(A_n) \end{aligned}$$

□

A.1.2 Bonferroni's Inequality

Statement: for fixed $n \in \mathbb{N}$, $P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - n + 1$.
Proof. Given the left hand side

$$\begin{aligned} P\left(\bigcap_{i=1}^{n} A_i\right) &= 1 - P\left(\overline{\bigcap_{i=1}^{n} A_i}\right) \\ &= 1 - P\left(\bigcup_{i=1}^{n} \overline{A_i}\right) \quad \text{By De Morgan's law} \\ &\geq 1 - \sum_{i} P(\overline{A_i}) \quad \text{By Boole's inequality} \\ &\geq 1 - \sum_{i} (1 - P(A_i)) \\ &\geq 1 - n + \sum_{i} P(A_i) \end{aligned}$$

□

A.2 Stochastic Convergence

A.2.1 Equivalent Conditions on almost sure convergence

A.2.2 Borel-Cantelli Lemma

A.2.3 Convergence almost surely implies convergence in probability

Proof. For simplicity, assuming the random variable $X_n \to 0$. Then by definition,

$$P\left(\limsup_{n\to\infty}\{|X_n| > \varepsilon\}\right) = 0$$

Then by Borel-Cantelli's lemma, almost sure convergence implies the convergence of the infinite sum of probabilities:

$$\sum_{n=1}^{\infty} P[\{|X_n| > \varepsilon\}] < \infty$$

Above implies that each term for sufficiently large n must converge to 0,

$$\lim_{n\to\infty} P[\{|X_n| > \varepsilon\}] = 0$$

We have X_n converges to 0 in probability. □

A.3 Non-existence of measure

Proof.

Construction:

Define the interval $I = (0, 1]$ with an equivalence relation $x \sim y$ if $x - y \in \mathbb{Q}$. That is:

$$[x] = \{x + r \mid r \in \mathbb{Q}, x \in I\}.$$

This partitions I into disjoint sets.

Pick $A \subseteq I$ with:

i) $\forall x, y \in A$, $x \sim y \implies x = y$,

ii) For each $x \in I$, if $x \in [x]$ for some x, then $x + r \in A$, where $A_i = x + A$.

Claim A.3.1:

Disjointness of Shifts: If $A_n = A + r_n$, where r_n is an enumeration of $\mathbb{Q} \cap (-1, 1)$, then:

$$A_n \cap A_m = \varnothing \quad \text{for } n \neq m.$$

Suppose $x \in A_n \cap A_m$. Then:

$$x \in A + r_n \quad \text{and} \quad x \in A + r_m.$$

This implies:

$$x = a + r_n \quad \text{and} \quad x = a' + r_m \implies r_n - r_m \in \mathbb{Q}.$$

By the construction of A, this forces $r_n = r_m$, which is a contradiction. Thus $A_n \cap A_m = \varnothing$ for $n \neq m$.

Claim A.3.2:

[Covering of $(0, 1]$]

$$(0, 1] \subseteq \bigcup_{n\in\mathbb{N}} A_n \subseteq (-1, 2).$$

(i) The first inclusion is given since all A_n serve as a partition of $(0, 1]$,

(ii) By construction, for all $x \in \bigcup_{n\in\mathbb{N}} A_n$, there exists $i \in \mathbb{N}$ such that $x \in A+r_i$. Since $A \subseteq (0, 1]$, we conclude:

$$x \in \mathbb{R} \cap (-1, 2).$$

By property (ii), $\mu(x + A) = \mu(A)$ for all $x \in \mathbb{R}$. By Claim 2:

$$\mu((0, 1]) \leq \mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) \leq \mu((-1, 2)).$$

We know $\mu((0, 1]) = C < \infty$. Then:

$$\mu((-1, 2)) = \mu((-1, 0]) + \mu((0, 1]) + \mu((1, 2]) = 3C.$$

Thus:

$$C \leq \sum_{n=1}^{\infty} \mu(A_n) \leq 3C.$$

If $\mu(A) > 0$, this leads to a contradiction such that an item bounded above by a finite value $3C$ diverges to infinity. Therefore:

$$\mu(A) = 0.$$

□

A.4 Proving Measurability

Proof 1. We prove this result in two directions.

Let $X : \Omega \to \mathbb{R}^d$ be a function. Then:

$$\sigma(X) = \sigma\left(X^{-1}(\mathcal{L})\right),$$

where $\mathcal{L}$ is a generator of the Borel σ-algebra $\mathscr{B}^d$.

1. Forward Direction: We show that:

$$X^{-1}(\mathcal{L}) \subseteq \sigma(X).$$

- By the definition of the σ-algebra $\sigma(X)$, it contains all preimages of measurable sets under X.
- Specifically, for all $A \in \mathscr{B}^d$, we know that $X^{-1}(A)$ is measurable.
- Since $\mathcal{L}$ is a generator of $\mathscr{B}^d$, any set in $\mathscr{B}^d$ can be written as a countable combination (union, intersection, complement) of sets in $\mathcal{L}$.
- Therefore, $X^{-1}(\mathcal{L}) \subseteq \sigma(X)$.

2. Reverse Direction: We show that:

$$\sigma(X) \subseteq \sigma\left(X^{-1}(\mathcal{L})\right).$$

- Define $\mathscr{F}_X = \{B \in \mathscr{B}^d \mid X^{-1}(B) \in \sigma(X^{-1}(\mathcal{L}))\} \subseteq \mathscr{B}^d$.
- We will skip proving the following two points:
  1. $\mathscr{F}_X$ is a σ-algebra.
  2. $\mathcal{L} \subseteq \mathscr{F}_X$ (principle of good sets).
- Since $\mathscr{B}^d$ is the smallest σ-algebra containing $\mathcal{L}$, we conclude that:

$$\mathscr{B}^d \subseteq \mathscr{F}_X.$$

- Therefore, for all $A \in \mathscr{B}^d$, $X^{-1}(A) \in \sigma(X^{-1}(\mathcal{L}))$.

Conclusion: Combining both directions, we obtain:

$$\sigma(X) = \sigma\left(X^{-1}(\mathcal{L})\right).$$

■

B Poisson Distribution from Binomial distribution

Theorem B.0.1

Let $\{X_n\}_{n\in\mathbb{N}^+}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{B}, P)$ and suppose that

$$X_n \sim \text{Binom}\left(n, \frac{\lambda}{n}\right)$$

Then $X_n \xrightarrow{d} X$ such that $X \sim \text{Poisson}(\lambda)$.

Proof. To show the convergence in distribution, we need to show that the probability densities converge to a limit density, that is,

$$\lim_{n\to\infty} P(X_n = k) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Fix $k \in \mathbb{N}^+$, given each $X_n \sim \text{Binom}\big(n, \frac{\lambda}{n}\big)$, we have

$$\begin{aligned}
\lim_{n\to\infty} P(X_n = k) &= \lim_{n\to\infty} \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k} && \text{definition of binomial coefficient} \\
&= \underbrace{\frac{\lambda^k}{k!}}_{\text{Poisson-ish}} \lim_{n\to\infty} \frac{n!}{(n-k)!}\left(1-\frac{\lambda}{n}\right)^{n}\frac{1}{n^k\big(1-\frac{\lambda}{n}\big)^k} && \text{grouping terms and factoring} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \{n(n-1)\cdots(n-k+1)\}\left(1-\frac{\lambda}{n}\right)^{n}\frac{1}{(n-\lambda)^k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \underbrace{\frac{\{n(n-1)\cdots(n-k+1)\}}{(n-\lambda)^k}}_{(1)}\underbrace{\left(1-\frac{\lambda}{n}\right)^{n}}_{\to e^{-\lambda}\text{ by def}}
\end{aligned}$$

Notice that term (1) expands to:

$$\begin{aligned}
\lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{(n-\lambda)^k} &= \lim_{n\to\infty}\frac{n^k}{n^k\big(1-\frac{\lambda}{n}\big)^k}\prod_{j=0}^{k-1}\left(1-\frac{j}{n}\right) \\
&= \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}\cdot\prod_{j=0}^{k-1}\left(1-\frac{j}{n}\right) \\
&= 1
\end{aligned}$$

The first term and each term in the product go to 1. Therefore, $\lim_{n\to\infty}(1) = 1$.

We have shown that $\exists X$ s.t. $\lim_{n\to\infty} X_n = X$ and $X \sim \text{Poisson}(\lambda)$. □

A more standard approach in proving weak convergence (convergence in distribution) is to look at the pointwise convergence in the characteristic functions. For $X_n \sim \text{Binomial}\big(n, \frac{\lambda}{n}\big)$,

Claim

$$\varphi_{X_n}(t) = \mathbb{E}\big[e^{itX_n}\big] = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^{it}\right)^n$$

Then the convergence becomes more apparent as

$$\begin{aligned}\lim_{n\to\infty} \varphi_{X_n}(t) &= \lim_{n\to\infty}\left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^{it}\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{\lambda(e^{it}-1)}{n}\right)^n \\ &= \exp\big\{\lambda\big(e^{it}-1\big)\big\}\end{aligned}$$

The result is precisely the characteristic function of a Poisson-distributed random variable.