MAT370 Intro to Mathematical Probability

Hanh H. Dao

December 23, 2025

Contents

Part I
Basic notations

A probability model consists of a sample space (which in turn defines a set of events), and a probability function, also known as a probability distribution.

1 Space

1.1 Measurable Space

Definition 1.1 (Measurable space)**.** A pair (X, Σ) is called a **measurable space** if X is a set and Σ is a nonempty σ-algebra of subsets of X.

1.1.1 Measure

Definition 1.2 (Measure μ)**.** Let (X, Σ) be a measurable space. A set function μ (inputs are sets) defined on Σ is called a **measure** if it satisfies the following properties:

1. $0 \leq \mu(A) \leq \infty$ for any $A \in \Sigma$.
2. $\mu(\varnothing) = 0$.
3. (σ**-additivity**) For any sequence of pairwise disjoint sets $\{A_i\}_{i=1}^{\infty} \subseteq \Sigma$, such that $\bigcup_{i=1}^{\infty} A_i \in \Sigma$, we have
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Remark. A measure input a set and output a number. That number can be the set's size, probability,...

1.2 Measure Space

Definition 1.3 (Measure Space and Probability Space)**.** A triplet (X, Σ, μ) is called a **measure space** if:

- (X, Σ) is a measurable space
- $\mu : \Sigma \to [0, \infty)$ is a measure.

If $\mu(X) = 1$, then μ is called a **probability measure**, usually denoted by $\mathcal{P}$, and the measure space is called a **probability space**.

1.3 Sample Space

Definition 1.4 (Sample space)**.** The sample space Ω is a set of all possible outcomes of a random experiment.

If the sample outcomes are denoted $\omega_1, \omega_2, \ldots$, then the sample space of an experiment can be

- a **finite** list of sample outcomes: $\{\omega_1, \ldots, \omega_k\}$,

- an **infinite** list of sample outcomes: $\{\omega_1, \omega_2, \ldots\}$,
- an **interval or region** of a real space: $\{\omega : \omega \in A \subseteq \mathbb{R}^d\}$.

Definition 1.5 (Probability Space)**.** A measure space (Ω, Σ, μ) is called a **probability space** if

$$\mu(\Omega) = 1.$$

In this case, μ is called a **probability measure**, usually denoted by P.

1.4 Probability Space

Definition 1.6 (Probability Space - General definition)**.** A measure space (Ω, Σ, μ) is called a **probability space** if

$$\mu(\Omega) = 1.$$

Definition 1.7 (Probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where:

- Ω is the sample space (set of outcomes ω),
- $\mathcal{F}$ is a σ-algebra of subsets of Ω (events),
- $P : \mathcal{F} \to [0, 1]$ is a probability measure with $P(\Omega) = 1$ and

$$P\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} P(E_j) \quad \text{for disjoint } E_j \in \mathcal{F}.$$

Definition 1.8 (Discrete probability space)**.** A discrete probability space includes:

- A sample space Ω that is finite or countably infinite.
- $\mathcal{F} = 2^{\Omega}$ (all subsets of Ω) as events.
- The probability measure P is defined by a function $p : \Omega \to [0, 1]$ with

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Then, for any event $E \subseteq \Omega$,

$$P(E) = \sum_{\omega \in E} p(\omega).$$

Intuition: An event is a sequence of many outcomes, then probability of that event is the sum of its outcomes.

2 Event

Definition 2.1 (Atomic ourcome)**.** An **atomic outcome** is a single member of the sample space.

Definition 2.2 (Sample outcome)**.** A sample outcome ω is precisely one of the (possible) outcomes of an experiment. Each $\omega \in \Omega$ is a general outcome (it can be a number, an event, a coin result, a sequence,...).

Definition 2.3 (Event)**.** An **event** is a subset of the sample space. Generally, an event is the collection of atomic outcomes.

3 Subsets and Power sets

Definition 3.1 (Power set)**.** The set of ALL subsets in A is called the power set of A, denoted $\mathcal{P}(A)$. The power set always contains A itself.

Theorem 3.2 (Size of a Power set)**.** If A is an n-element set, then $\mathcal{P}(A)$ has 2^n elements. In other words, an n-element set has 2^n distinct subsets.

Definition 3.3. Event A occurs if the outcome of the random experiment is a member of the set A.

Definition 3.4 (Support)**.** Suppose that $f : X \to \mathbb{R}$ is a real-valued function whose domain is an arbitrary set X. The **set-theoretic support** of f, written $\operatorname{supp}(f)$, is the set of points in X where f is non-zero:

$$\operatorname{supp}(f) = \{\, x \in X : f(x) \neq 0 \,\}.$$

4 Sigma-algebra

Definition 4.1 (Sigma-algebra)**.** A sigma-algebra (σ-algebra or σ-field) $\mathcal{F}$ is a set of subsets ω (events) of sample space Ω such that:

1. The empty event is always measurable: $\varnothing \in \mathcal{F}$.
2. $\mathcal{F}$ is closed under **complements**: If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. $\mathcal{F}$ is closed under **countably finite unions**: If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Remark. Sigma algebras can be generated from arbitrary sets. This will be useful in developing the probability space.

Remark. The power set $\mathcal{P}(\Omega)$ does contain ALL subsets of Ω but the σ-algebra generated by Ω, denoted as σ, may contain some subsets (=events that can occur).

Definition 4.2. The σ-algebra generated by Ω, denoted Σ, is the collection of possible events from the experiment at hand.

Example 1. Consider an experiment with $\Omega = \{1, 2\}$. Then, the σ-algebra is

$$\Sigma = \{\varnothing, \{1\}, \{2\}, \{1, 2\}\}.$$

Part II
Counting

Example 2 (Drawing balls from an urn)**.** When drawing balls from an urn of m distinguishable balls (they are labeled 1,2,...,m), we wanna count how many outcomes:

1. **Sampling**: drawing objects under conditions)
2. **Allocating**: distributing objects into boxes

5 Sampling (permutations and combinations)

Every combinatorial "sampling" problem can be classified by two choices:

1. Replacement
2. Ordering

5.1 With replacement + with ordering

Setup 5.1.

- An urn with m distinguished balls
- Draw the balls n times, after each draw, we put the ball back.
- We care about the order in which balls appear.

Formula 5.2. Apply the Fundamental rule:

$$m^n$$

1. Each outcome is an **ordered n-tuple** $(a_1, a_2, \ldots, a_n)$ of drawn balls.
2. Each $a_j, 1 \leq j \leq n$ can be any of the m balls.
3. m^n = the number of ordered n-tuples with repetition.

5.2 With replacement + without ordering

Setup 5.3.

- An urn with balls of m different colors.
- Draw the balls n times, with replacement (color can appear multiple times), and without order (only the counts of colors matter).

Formula 5.4.

$$\binom{m+n-1}{n} = \binom{m+n-1}{m-1}$$

1. Each outcome is a **multi-set** of size n from a set of size m.
2. $\binom{m+n-1}{m-1}$ = number of ways to arrange n checks and $m-1$ bars = number of multisets of size n drawn from m types.

Case 3: Without replacement + with ordering (Permutation)

$$ {}_mP_n = \frac{m!}{(m-n)!} = (m)_n $$

- We choose n balls and care about the order.
- Special case (all m balls arranged in order): when $m = n$,

$$ {}_mP_m = \frac{m!}{(m-m)!} = m! $$

Case 4: Without replacement + without ordering (Combination)

$$ {}_mC_n = \binom{m}{n} = \binom{m}{n} = \frac{m!}{n!(m-n)!} $$

- We choose n balls out of m, and order does not matter.
- Called the **binomial coefficient**.
- Orders of k chosen and $n-k$ unchosen do not matter

Formula 5.5 (Permutation and Combination)**.**

$$ P(n,r) = \binom{m}{n} \cdot n! $$

- Combination = Which objects? (choose n from m)
- Permutation = Which objects, and in what order? (arrange n in all possible orders)

6 Binomial formula

Remark. The binomial formula is just a compact expression of:

1. **Probability of one pattern**:
$$ p^k(1-p)^{n-k} $$
2. **Count all patterns**:
$$ \binom{n}{k} p^k(1-p)^{n-k} $$

- INDEPENDENT events: sampled with replacement: use distribution types
- DEPENDENT events: sampled without replacement: use counting/combinatorial/hypergeometric methods

7 Multinomial Coefficient

The Multinomial coefficient is number of ways to permute m objects that are distinguishable by groups.

Definition 7.1 (Multinomial coefficient)**.**

$$ \frac{m!}{m_1!\, m_2! \cdots m_n!} $$

This is number of distinct arrangements to partition all m balls into n labeled groups. Groups can have size $m_1, m_2, \ldots, m_n$.

- Special case: when $n = 2$, multinomial coefficient becomes a binomial coefficient:

$$\binom{m}{m_1} = \binom{m}{m_2}$$

.

8 Examples

8.1 Symmetry

Example 3 (1.12 - Flipping coins)**.** Suppose we flip seven coins. Compute the probability that we get 0, 1, 2, or 3 heads.

Solution 3.1. Flipping the coins is a type of combination.

1. Total outcomes $= 2^7 = 128$.
2. 1 head: There are 7 outcomes : ${}_7C_1 = \binom{7}{1} = 7$
3. 2 heads, 3 heads, ...: ${}_7C_2 = \binom{7}{2}, {}_7C_3 = \binom{7}{3}, \ldots$
4. By symmetry, $\binom{7}{1} = \binom{7}{6}, \binom{7}{5} = \binom{7}{2}, \binom{7}{4} = \binom{7}{3}$

Property 9 (Symmetry)**.** By words, If you want to choose m objects to take, that's the same as saying you've chosen $n - m$ objects to leave behind.

$$\binom{n}{m} = \binom{n}{n - meco}$$

9.1 Comparing mathematical models with real-world data

Example 4 (1.13 - World Series)**.** We model a game tournament as:

- Two teams play a best-of-7 series.
- The first team to win 4 games wins the championship.
- So the series can last 4, 5, 6, or 7 games (never less, never more).

Assume that each team is equally strong. A wins with probability of 1/2 and so does B. All games are independent. What are the probabilities of the series lasting 4,5,6, or 7 games?

Solution 4.1. The series length depends on when the 4th win happens.

- Total outcome: $2^4 = 16$
- 4 games: one team (A and B) wins all 4 games (AAAA or BBBB) so $P(A) = 2/16$.
- 5 games: The 5th win must be either A's or B's. In the first 4 games, there are exactly 3 wins and the number of sequence like this is $\binom{4}{3} = 4$. By symmetry, there are 8 ways and $P(5) = 4/16$.
- Similar things to 6 and 7 games.

So the longer the series, the more possible sequences allow it to happen $\rightarrow$ the higher the probability. So 6 or 7 games are most likely.

Remark. The real-world data shows deviations $\rightarrow$ meaning the fair-coin assumption is not perfect. So the example illustrates both combinatorial probability and the importance of comparing models with data.

9.2 Combinatorics and large number of pairs

Example 5 (4.17 - Birthday problem)**.** There are 30 people at a birthday party. What is the chance that there is at least 2 people have the same birthday?

Part III
Random Variable

Definition 9.1. Given a sample space Ω, a random variable is a function $X : \Omega \to \mathbb{R}$. Notation: $X(\omega)$ is a real number with $\omega in \Omega$.

10 Distribution of a Random Variable

Lemma 10.1 (Distribution of a RV)**.** A random variable X and probability measure $\mathcal{P}$ induce a probability measure μ on Borel sets $B \subseteq \mathbb{R}$ defined by:

$$\mu_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

This measure μ_X is the **distribution** of X.

Observation 10.2. Differentiate X and μ_X:

1. $P(\{X \in B\}$ is the probability on Ω
2. μ_X is the probability on $\mathbb{R}$, give number between 0 and 1
3. X is a function that outputs to any real nnumbers

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

11 Types of Random Variables

- Discrete: μ_X gives positive probability only to countable points.
- Continuous: $\mu_X(\{x\}) = 0$ for every real x.
- General: every random variable can be split into a discrete part + continuous part.

11.1 Discrete Random Variable

Definition 11.1 (Discrete RV)**.** A random variable X is called a discrete random variable if its range (the set of possible values) is finite or countably infinite

12 Distribution Functions

12.1 Cummulative Distribution Function - CDF

Instead of listing probabilities of all possible outputs of a variable, we use a function F_X.

Definition 12.1 (Cummulative distribution function)**.**

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

12.1.1 Properties of CDF

Theorem 12.2. Three main properties of $F_X(x)$:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
2. F is a nondecreasing function.
3. F is right continuous: $\lim_{y\to x^+} F(y) = F(x)$. Even if there is a jump from the left, the value at the point is determined from the rights.

Theorem 12.3. If a function $F : \mathbb{R} \longrightarrow [0, 1]$ satisfies the three properties above, then F is the CDF of some random variable.

Theorem 12.4. A distribution function F for a random variable X uniquely defines the measure μ_X.

12.1.2 CDF for Discrete RV

Formula 12.5. CDF for any discrete random variable with pmf $p(x)$:

$$P(X \le k) = \sum_{x\le k} p(x)$$

12.1.3 CDF for Continuous RV

Formula 12.6. CDF for any continuous random variable with pdf $f(x)$:

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

12.2 Probability Density Function - PDF

12.3 Probability Mass Function - PMF

Definition 12.7 (Probability Mass Function)**.** The distribution F_X of a discrete random variable X is determined by its PMF:

$$p_X(x) = P(X = x)$$

Property 13 (PMF properties)**.**

1. Non-negativity: $\forall x,\ p_X(x) \ge 0$
2. Normalization: $\sum_{x\in \text{Range}(X)} p_X(x) = P(X \in \mathbb{R}) = 1$

14 Mathematical Expectation

Definition 14.1 (Expectation)**.** Expectation = weighted average of values of X.

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)P(\{\omega\}),$$

provided the absolute convergence condition holds:

$$\sum_{\omega\in\Omega} |X(\omega)|P(\{\omega\}) < \infty.$$

That condition ensures the sum makes sense (doesn't blow up or oscillate).

15 General Random Variables

But there exist "mixed-type" random variables and singular distributions that do not fit into either "discrete" or "continuous" categories.]

Example 6 (Cantor distribution)**.** In mathematics, the Cantor function is an example of a function that is continuous, but not absolutely continuous. It is a notorious counterexample in analysis, because it challenges naive intuitions about continuity, derivative, and measure

The *Cantor function* $C : [0, 1] \to [0, 1]$ is defined as follows. For any $x \in [0, 1]$:

1. Express x in base 3, using digits $0, 1, 2$.
2. If the base-3 representation of x contains a 1, replace every digit strictly after the first 1 with 0.
3. Replace any remaining 2s with 1s.
4. Interpret the resulting sequence as a binary number. The result is $C(x)$.

Remark. $C(x)$ is continuous, increasing, and flat on many intervals.

16 Functions of Random Variable

Theorem 16.1.

$$f_Y(y) = f_X\big(s(y)\big)\,|s'(y)|.$$

If r is increasing, then $s'(y) > 0$ and $|s'(y)| = s'(y)$.
If r is decreasing, then $s'(y) < 0$ and $|s'(y)| = -s'(y)$.

Part IV
Expectation

17 Expectation

Definition 17.1. Expectation is a probability-weighted average of all possible values, and the

Let X be a random variable.

- **If X is discrete** with PMF $P(X = x)$, its expectation (or mean) is defined by
$$\mathbb{E}[X] := \sum_x x \cdot P(X = x),$$
where the sum is over the range of X.
- **If X is continuous** with PDF $f(x)$, its expectation (or mean) is defined by
$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

17.1 Expectation of a Product

Formula 17.2. We must use joint distribution to compute epectation of a product:

1. Discrete case:
$$E[XY] = \sum_x \sum_y xy\, P(X = x, Y = y)$$
2. Continuous case:
$$E[XY] = \iint xy\, f_{X,Y}(x, y)\, dx\, dy$$

Lemma 17.3. If X and Y are independent, then $E(XY) = E(X)E(Y) = f_X f_Y$.

18 Survival function

Definition 18.1. $P(X > x)$ is called the survival function — it tells you how likely the variable exceeds x.

- You dont need to know the pdf $f(x)$, but you can estimate tail probabilities.
- Large values of X, even if rare, contribute significantly to the expectation.

18.1 Expectation via Tail

Lemma 18.2 (Expectation via tail)**.** Let X be a non-negative, integer-valued random variable. Then,
$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

19 Median

Definition 19.1. Let X be a random variable. A number $m_{1/2} \in \mathbb{R}$ is called a *median* of X if

$$P(X \leq m_{1/2}) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq m_{1/2}) \geq \frac{1}{2}.$$

20 Partial expectation

The expectation of $f(x)$ but only when event A happens.

$$E(f(X) \cdot \mathbf{1}_A)$$

- It is the plain expectation over the whole probability space, but only when A occurs.

Remark. It is NOT the same as $f(X) \cap A$ which is meaningless: one is number amd the other is set/event.

21 Conditional expectation

Average value of $f(X)$ given that we know A happened:

$$E(f(X) \mid A) = \frac{E(f(X)\mathbf{1}_A)}{P(A)}, \qquad \text{provided } P(A) > 0.$$

Remark. Comparisons:

- $\mathbf{1}_A$ = filter
- $|A$ = average after filtering

Part V
Variance

22 Variance

Definition 22.1. Let X be a random variable. The variance of X is defined by

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

Lemma 22.2. Let X be a random variable whose variance exists, then:

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

23 Properties

1. Variance is not a linear operator: $\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2 \operatorname{Cov}(X, Y)$
2. The variance of the sum equals all variances plus all pairwise covariances.

$$Var(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n} Var(X_i) + \sum_{1 \le i,j \le n} Cov(X_i, X_j).$$

Remark.

Lemma 23.1. If X and Y are independent, then $Cov(X, Y) = 0$ and,

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

24 Covariance

Definition 24.1. Let X, Y be random variables. The covariance of X and Y is defined by

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

If $\operatorname{Cov}(X, Y) = 0$, we say that X and Y are *uncorrelated*. Uncorrelated is NOT independent.

24.1 Independence

Formula 24.2. A true identity:

$$E[XY] = \operatorname{Cov}(X, Y) + E[X]E[Y]$$

- If covariance $= 0$, then $E[XY] = E[X]E[Y]$ (but this does not imply independence in general)
- If covariance $\neq 0$, it adds or subtracts value.

Remark. Expectation of a Product always involves the joint distribution.

25 Cauchy–Schwarz In equality

Theorem 25.1 (Cauchy–Schwarz for random variables)**.**

$$|E[XY]| \leq E[|XY|] \leq \sqrt{E\left[X^2\right]}\sqrt{E\left[Y^2\right]}$$

where $\langle X, Y\rangle = E[XY]$ and $\|X\| = \sqrt{E[X^2]}$.

- Geometric: $|\langle X, Y\rangle| \leq \|X\|\,\|Y\|$ (inner product of two vectors)
- $E[XY]$ measures how much X and Y "move together": $\mathrm{Cov}(X, Y) = E\left[(X - E[X])\,(Y - E[Y])\right]$

Remark. Two random variables can align (their covariance-type term) is fundamentally limited by their individual sizes (variances).

Part VI
Distribution of Discrete RV

In probability theory, random variable can take on any real values, depending on the distribution.

26 Bernoulli

In a Bernoulli trial, you perform one experiment (e.g., tossing a coin that lands heads with probability p).
The Bernoulli random variable X represents the outcome of that single trial. The outcome is binary.

$$X = \begin{cases} 1, & \text{if success occurs (e.g., head),} \\ 0, & \text{if failure occurs (e.g., tail).} \end{cases}$$

Definition 26.1. $X \sim \text{Bernoulli}(p)$ if

$$P(X = x) = \begin{cases} p, & \text{if } x = 1, \\ 1 - p, & \text{if } x = 0. \end{cases}$$

Expectation:

$$\mathbb{E}[X] = p.$$

Variance:

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

27 Geometric

Definition 27.1 (Geometric distribution)**.**

$$P(X = k) = (1 - p)^{k-1} p$$

- Number of trials until 1st success: k
- Probability of success for each trial: p

1. Memoryless property:

$$P(X = m + n \mid X > m) = P(X = n) \quad \text{or } X \mid (X > m) \stackrel{d}{=} m + X$$

2. Tail probability:

$$P(X > k) = (1 - q)^k$$

3. Another form: $P(Y = k) = P(Y > k - 1) - P(Y > k) = (1 - q)^{k-1} q$.

27.1 Geometric series

Definition 27.2. Sum of geometric series:

1. Finite Geometric Sum:
$$\sum_{k=0}^{n-1} ar^k = a\frac{1-r^n}{1-r}, \qquad r \neq 1.$$

2. Infinite Geometric Sum:
$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}, \qquad |r| < 1.$$

27.2 Expectation

1. Expectation: $E(X) = \sum_{k=1}^{\infty} kP(X=k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p \cdot \frac{1}{p^2} = \frac{1}{p}$

27.3 Variance

Formula 27.3. Variance of X where $X \sim \text{Geo}(p)$ is:

$$\text{Var}(X) = \frac{1-p}{p^2}.$$

Proof. Sketch proof with Sum of Infinite Geometric series:

- Start from the geometric series identity
$$\sum_{n=0}^{\infty} x^n = (1-x)^{-1}.$$
- Differentiate once and twice to obtain formulas for $\sum nx^{n-1}$ and $\sum n(n-1)x^{n-2}$.
- Substitute $x = 1-p$ so the powers match the geometric pmf.
- Multiply by $p(1-p)$ to obtain
$$E[N(N-1)].$$
- Use the identity
$$E[N^2] = E[N(N-1)] + E[N], \qquad E[N] = \frac{1}{p}.$$
- Compute
$$\text{Var}(N) = E[N^2] - (E[N])^2 = \frac{1-p}{p^2}.$$

□

28 Hyper-geometric

Definition 28.1 (Hypergeometric Distribution)**.**

$$P(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, \ldots, \min(m,n).$$

Probability of selecting k specials (**without replacement**) from a sample size n, where there are m "special" items in population N.

28.1 Components

For each draw i,

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th draw is red,} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\mathbf{X} = X_1 + \cdots + X_k.$$

Each X_i is Bernoulli, but the X_i are not independent as we draw **without replacement**.

- Mean: kp
- Covariance: the indicators X_i are negatively correlated.

$$\operatorname{cov}(X_i, X_j) = -\frac{p(1-p)}{n-1}.$$

- Variance:

$$\operatorname{var}(R_k) = k\,p(1-p)\left(1 - \frac{k-1}{n-1}\right).$$

  - $k\,p(1-p)$: the binomial variance.
  - $\left(1 - \frac{k-1}{n-1}\right)$: the finite population correction.

28.1.1 Hyper-geometric and Binomial

When the population is large and the sample is small,

$$\text{Hypergeometric}(n, m, k) \approx \text{Binomial}\left(k, \frac{m}{n}\right).$$

Then the means match:

$$kp \approx k \cdot \frac{m}{n}.$$

29 Multinomial

Definition 29.1 (Multi-nominal Distribution)**.** A vector $X = (X_1, X-2, \cdots, X_n)$ has a multinominal distribition with joint pmf:

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

We repeat the same experiment n times. Each trial can result in one of k categories. Let X_i denote the number of times category i occurs, for $i = 1, 2, \ldots, k$, with

$$X_1 + X_2 + \cdots + X_k = n.$$

1. Marginal Distribution becomes Binomial:

$$P(X_1 = n_1) = \binom{n}{n_1} (p_1)^{n_1} (1-p_1)^{n-n_1}$$

Remark. In general, if each of n independent trials results **(with replacement)** in one of k categories with probabilities $p_1, p_2, \ldots, p_k$, then the counting vector $X = (X_1, X_2, \ldots, X_k)$ follows a Multinomial$(n; p_1, \ldots, p_k)$ distribution with PMF:

$$P(X_1 = n_1, \ldots, X_k = n_k) = \binom{n}{n_1, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

whenever $n_i \geq 0$ and $n = n_1 + n_2 + \cdots + n_k$.

Remark. Though multi-nomial has independent trials, the category counts are dependent (sum of them has to add up to n). Hence we multiply their fized probabilities of each possible counts.

30 Poisson

Definition 30.1 (Poisson Distribution)**.** $X \sim \text{Poisson}(\lambda_0)$ with $k = 1, 2, \cdots$

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

- X = number of events in a fixed time
- λ = average number of events per unit of time
- k = exact number of events you observe

Remark. We only count how many events. The average number of events is $\lambda = np$

Theorem 30.2 (Poisson Approximation)**.** If

$$S_n \sim \text{Binomial}(n, p_n), \qquad p_n \to 0, \qquad np_n \to \lambda,$$

then

$$S_n \approx \text{Poisson}(\lambda).$$

Part VII
Continuous RV - Distributions

31 Exponential

If $X \sim \text{Exp}(\lambda)$ for some $\lambda > 0$,

1. Interpretation:
   - Variable X models waiting time until an event occurs.
   - Rate parameter λ tells how quickly the probability decays as x increases. Large λ means events happen quickly, small average waiting time and conversely.
2. Support: $[0, \infty)$
3. PDF:
$$f(x) = \lambda e^{-\lambda x}$$
4. CDF:
$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$
5. Median: $P(X \leq m_{\frac{1}{2}}) = 0.5 \implies m_{\frac{1}{2}} = \frac{\ln 2}{\lambda}$
6. Expectation $\sim$ expected waiting time until an event:
$$\mathbb{E}[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}$$

32 Uniform

33 Normal

Theorem 33.1. CLT: A sum of many independent (or weakly dependent) random variables with finite mean and variance is approximately normal.

- Sum of Bernoulli trials
- Sum of Poisson variables with Poisson(1)

34 Hierarchy

Distribution approximation hierarchy

$$\text{Hypergeometric} \longrightarrow \text{Binomial} \longrightarrow \text{Poisson}.$$

1. Hypergeometric: exact sampling without replacement (finite population).
2. Binomial: approx. when N large and n is small, pretends sampling with replacement (constant p)
3. Poisson: approx. when n is large, p is small, and $np = \lambda$ (rare events).

Part VIII
Conditional Probability

35 Conditional Probability

Definition 35.1. The probability of event B occurring given that event A has already occurred is defined as:

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

35.1 Multiplication rule

Formula 35.2. From joint probability to conditional probability:

$$P(A) \cdot P(B|A) = P(B \cap A)$$

35.2 Chain Rule for Conditioning

Theorem 35.3. To compute the probability that several events all happen $P(A_1, \ldots, A_n)$:

$$P(A_1, \ldots, A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \cdots P(A_n \mid A_1, \ldots, A_{n-1})$$

Each new event is conditioned on all the previous events happening.

35.3 Properties of Conditional probability

Severla properties of Conditional probability:

1. **Range:**
$$0 \leq P(B \mid A) \leq 1$$
The probability is always between 0 and 1.
2. **Normalization:**
$$P(\Omega \mid A) = 1$$
where Ω represents the sample space (since Ω always occurs).
3. **Additivity:** If events B_i are disjoint, then the probability of their union, given A, is:
$$P\left(\bigcup_i B_i \mid A\right) = \sum_i P(B_i \mid A)$$
This is similar to how probabilities work for ordinary events.
4. **Complement Rule:** For the complement of an event B, we have:
$$P(B^c \mid A) = 1 - P(B \mid A)$$
5. **Disjoint $\neq$ Independence**: A and B are disjoint events that have positive probability, then they are not independent since
$$P(A)P(B) > 0 = P(A \cap B)$$

Remark. $P(P|C^c) \neq 1 - P(P^c|C^c)$

36 Bayes's Formula

Formula 36.1. Bayes' Rule updates **what we believe** about an event B_j after observing some **evidence** A:

$$P(B_1 \mid A) = \frac{P(B_1)\,P(A \mid B_1)}{\sum_{i=1}^{n} P(B_i)\,P(A \mid B_i)} = \frac{P(B_1 \cap A)}{\sum_{i=1}^{n} P(B_i \cap A)} = \frac{P(B_1)\,P(A \mid B_1)}{P(A)}$$

- Prior probability $P(B_1)$: what we believed about B_1 before seeing evidence A
- Likelihood $P(A|B_1)$: probability of observing A assuming B_1 is true
- Evidence $P(A)$

Posterior = Prior $\times$ Likelihood $\div$ Evidence

36.1 Conditional and Bayes's Rule

Conditional probability:

- Conditional probability: "Inside A, what fraction is B?"
- Bayes' rule: "Inside A, how is probability split across all possible B_i?"

36.2 Maximum A Posteriori Estimate: MAP

Definition 36.2. Given:

- A parameter or hypothesis θ (coin type, model, class label, etc.)
- Observed data x

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta \mid x) = arg\max_{\theta}[P(x \mid \theta)P(\theta)]$$

MAP is a way to choose the parameter (or model) that is most probable after seeing the data, using Bayes' rule.

37 Law of Total probability

Formula 37.1 (Law of Total probability)**.**

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

Formula 37.2. If $\{B_1, B_2, \ldots, B_n\}$ is a partition of conditional event C (mutually exclusive and exhaustive), then for any event A conditioned on C, $P(A) = \sum_{k=1}^{n} P(A \cap B_k)$.

$$P(A \mid C) = \sum_{k=1}^{n} P(A \cap B_k \mid C) = \sum_{k=1}^{n} P(A \mid B_k, C)\,P(B_k \mid C).$$

Remark. Forward probability: "What is the overall chance of seeing event A, considering all possible reasons B_i that could cause it?

38 Two-stage experiment

Classic structure:

1. Stage 1: Partition the 1st stage into exclusive events.
2. Stage 2: Use the multiplication rule to compute joint events (stage 1 + stage 2). Add them = law of total probability.

At each stage total probability is 1.

39 Examples

39.1 Example 5.3 - Bridge

- There are 4 players: North, South, West, and East.
- Each player is dealt 13 cards from a deck of 52 cards.
- North and South together have 8 Hearts.
- What is the probability of West have 3 hearts and 2 non-hearts?

Solution 6.1. North and South together have 8 Hearts $\implies$ there are left with 5 hearts, and 21 non-hearts card. P(3 hearts) = $\frac{\binom{5}{3}\cdot\binom{21}{10}}{\binom{26}{13}}$

Part IX
Independence

40 Inclusion-exclusion principle

Formula 40.1 (**Poincaré's Formula**)**.** For arbitrary events $A_1, \ldots, A_n$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) + \sum_{1\le i<j<k\le n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

where the indices in each sum are distinct and range from 1 to n.

- Inclusion-exclusion principle (n=2):

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

41 Independence

Definition 41.1 (Independence)**.** Events A and B are independent **if and only if** either

$$P(B \mid A) = P(B)$$

(or equivalently, $P(A \mid B) = P(A)$).
or

$$P(B \cap A) = P(A)P(B)$$

This means: knowing that A happened does not change the probability that B happens. Those 2 statements are mathematically equivalent.

Remark. With no assumption, $P(B \cap A) = P(A \mid B)P(B)$ is ALWAYS USED. Only use $P(B \cap A) = P(A)P(B)$ when we already know A and B are independent.

42 Multiple events

Situation	Probability wanted	Formula
Either A or B (general)	$P(A \cup B)$	$P(A) + P(B) - P(A \cap B)$
Either A or B (exclusive) - cannot occur at the same time	$P(A \cup B)$	$P(A) + P(B)$
Both A and B independent	$P(A \cap B)$	$P(A) \cdot P(B)$
Both A and B dependent	$P(A \cap B)$	$P(A) \cdot P(B \mid A)$

42.1 Pairwise independence

Definition 42.1 (Pairwise independence)**.** Events $A_1, \ldots, A_n$ are *pairwise independent* if

$$P(A_i \cap A_j) = P(A_i)\, P(A_j), \qquad \text{for all } i \neq j.$$

Each pair of events is independent, but groups of 3 or more may not be.

42.2 Full independence

Definition 42.2 (Full independence)**.** Events $A_1, \ldots, A_n$ are *independent* if for any $1 \leq i_1 < \cdots < i_k \leq n$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}).$$

All subsets of events behave independently (not just pairs).

43 Mutually exclusive vs Independence

1. Mutually exclusive (disjoint events): cannot happen at the same time $A \cap B = \varnothing$

$$P(A \cap B) = 0$$

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

2. Indenpendent events:

$$P(B \mid A) = P(B)$$

or

$$P(A \cap B) = P(A)\, P(B)$$

Part X
Sums of Independent Random Variables

44 Discrete case

(General case)	$P(X+Y=z) = \sum_x P\big(X=x,\ Y=z-x\big),$
(Independent case)	$P(X+Y=z) = \sum_x P(X=x)\, P\big(Y=z-x\big).$

44.1 Sum of Binomial variables

Definition 44.1. Sum of binomial variables is a binomial variable:

$$\boxed{S \sim \text{Binomial}(n_1+n_2+\ldots+k_k,\ p)}$$

Formula 44.2 (Binomial Sum Distribution)**.**

$$P(S=k) = \binom{N}{k} p^k (1-p)^{N-k}, \qquad S \sim \text{Binomial}(N,p).$$

- Model $S=k$: total number of successes
- N: total trials
- p: probability of success

44.2 Sum of Poisson variables

Definition 44.3. If

$$X_i \sim \text{Pois}(\lambda_i), \qquad i=1,\ldots,n,$$

are If

$$X_i \sim \text{Pois}(\lambda_i), \qquad i=1,\ldots,n,$$

are **independent**, then

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

, then

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

44.3 Sum of Geometric variables = Negative Binomial Distribution

$$X_1,\ldots,X_n \sim \text{i.i.d. Geometric}(p)$$

$$S = \sum_{i=1}^n X_i$$

.

- S is number of trials needed to see n successes.
- Each X_i is the number of trials until i-th success.
- This distribution is called Negative Binomial.

Definition 44.4 (Negative Binomial Distribution)**.**

$$P(S = m) = \binom{m-1}{n-1} p^n (1-p)^{m-n}, \qquad T_n \sim \text{NegBin}(n, p).$$

NegBin(n, p) models how many trials you wait until n successes, when each trial succeeds with probability p.

Example 7. To prove the formula, consider the following outcome with $m = 10$ and $n = 3$:

$$\text{FSFFFFSFFS}.$$

Any string with 3 successes (S) and 7 failures (F) has probability

$$p^3(1-p)^7.$$

For $T_3 = 10$, the last letter must be S. The other $n-1$ locations for S are chosen from the first $m-1$ positions.

Part XI
Limiting Theorem

45 Sequence of Random Variables

- Sample space: $S = \{s_1, s_2, \ldots, s_k\}$
- Random variable:
$$X(s_i) = x_i, \quad \text{for } i = 1, 2, \ldots, k.$$
- Sequence of random variables:
$$X_n(s_i) = x_{ni}, \quad \text{for } i = 1, 2, \ldots, k.$$
- Sequence of random variables is a sequence of function $X_n : S \longrightarrow \mathbb{R}$

Definition 45.1. X_n is random variable defined on **the same experiment**, but **its mapping changes with** n.

46 Indicator Random Variable

Definition 46.1 (Indicator variable)**.** An indicator random variable I is a special kind of random variable based on such an event, and it equals 1 if the event happens and 0 otherwise.

Formula 46.2. For an event $E \subseteq \Omega$, the indicator function is:
$$1_E(\omega) = \begin{cases} 1, & \omega \in E, \\ 0, & \omega \notin E. \end{cases}$$

46.1 Expectation of Indicator RV

Formula 46.3. For any indicator I,
$$\mathbb{E}(I) = 0 \cdot P(I = 0) + 1 \cdot P(I = 1) = P(I = 1) = p$$

Formula 46.4 (General identity)**.**
$$E\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} P(\text{event}_i)$$

46.2 Identity

Formula 46.5. For any random variable X and an event A,
$$E[X \cdot \mathbf{1}_A] = P(A)E[X \mid A]$$

- $\mathbf{1}_A$: indicator of event A
- X: any random variable
- $E[X \mid A]$: conditional expectation given event A

- A is an event (a condition), and X is a random variable defined on the same probability space.

Remark. Interpretation: Probability of A $\times$ average value of X when A happens.

Formula 46.6 (General split with an event A)**.** For any integrable random variable X,

$$E[X] = E[X \cdot \mathbf{1}_A] + E[X \cdot \mathbf{1}_{A^c}]$$

Equivalent Law of Expectation form:

$$E[X] = P(A)\,E[X \mid A] + P(A^c)\,E[X \mid A^c]$$

Countable partition version:

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} E[X \cdot \mathbf{1}_{A_i}] \\ &= \sum_{i=1}^{\infty} P(A_i)\,E[X \mid A_i]. \end{aligned}$$

This assumes that $\{A_i\}_{i\geq 1}$ is a disjoint partition of the sample space Ω, meaning $\bigcup_{i=1}^{\infty} A_i = \Omega$.

46.3 Bernoulli Variable

Definition 46.7. Every indicator variable is a Bernoulli random variable.

Formula 46.8.

$$P(X = x) = \begin{cases} p, & \text{if } x = 1, \\ 1 - p, & \text{if } x = 0. \end{cases}$$

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

Aspect	Bernoulli RV	Indicator variable
Viewpoint	Probabilistic	Logical
Defined by	Distribution	Event
Purpose	Modeling randomness	Counting events
Typical use	Single-trial model	Proofs, expectations, sums

47 Central Limit Theorem

Definition 47.1. If $X_1, X_2, \ldots$ are **independent and identically distributed** with mean μ and variance σ^2, then for all real numbers $a < b$,

$$\mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx, \qquad \text{as } n \to \infty.$$

$\frac{S_n - n\mu}{\sigma\sqrt{n}}$ is a random variable whose convergence in distribution to standard normal distribution.

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

47.1 Standardization

Standardized sum:

1. Starts with a sequence of i.i.d. random variables:

$$X_1, X_2, X_3, \ldots$$

2. Then, for each n, it defines a new random variable

$$S_n = X_1 + X_2 + \cdots + X_n.$$

3. We have a sequence:

$$S_1, S_2, S_3, \ldots$$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

This is the object to which the Central Limit Theorem applies.

47.2 Continuity correction

- Discrete distributions $\to$ bars.
- Continuous distributions $\to$ curves.
- Correction $[\,k - 0.5,\ k + 0.5\,]$.

47.3 CLT and Chebyshev's Inequality

48 Convergence in Distribution

Definition 48.1. Given:

- A sequence of random variables $X_1, X_2, \ldots$
- A corresponding sequence of cdfs $F_{X_1}, F_{X_2}, \ldots$ so that for $n = 1, 2, \ldots,$

$$F_{X_n}(x) = P(X_n \leq x).$$

Suppose that there exists a cdf F_X such that for all x at which F_X is continuous,

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

Then

1. F_X is the limiting distribution.
2. $X_1, \ldots, X_n$ converges in distribution to random variable X with cdf F_X, denoted

$$X_n \xrightarrow{d} X,$$

Convergence of a sequence of mgfs or cfs also indicates convergence in distribution, that is, if for all t at which $M_X(t)$ is defined, as $n \to \infty$ we have

$$M_{X_n}(t) \to M_X(t) \quad \iff \quad X_n \xrightarrow{d} X.$$

49 Convergence in Probability

Definition 49.1. X_n converges to θ in probability, written $X_n \xrightarrow{p} \theta$, if, for every $\epsilon > 0$,

$$P(|X_n - \theta| > \epsilon) \to 0.$$

Proof. If $\text{Var}(X_n) \to 0$, then

$$\frac{\text{Var}(X_n)}{\epsilon^2} \to 0$$

So

$$P(|X_n - \mathbb{E}[X_n]| > \epsilon) \to 0$$

which means

$$X_n \xrightarrow{P} \mathbb{E}[X_n]$$

And if $\mathbb{E}[X_n] \to \theta$, then

$$X_n \xrightarrow{P} \theta$$

□

Remark (Intuition)**.** Given:

- Data: $X_1, X_2, \ldots, X_n$, each X_i is an observation
- True parameter θ (an unknown constant)
- Estimator computed from data: $\hat{\theta_n}$ (random, as it depends on random sample)

$$\hat{\theta}_n \xrightarrow{P} \theta$$

As n grows, the estimator $\hat{\theta}_n$ gets closer to the true parameter θ with probability approaching 1.

Remark. Comparisons of 2 types of Convergence:

- **In probability:** $P(|X_n - X| > \epsilon) \to 0$ for all $\epsilon > 0$
- **In distribution:** $F_n(t) \to F(t)$ at continuity points t

49.1 Markov's inequality

Theorem 49.2. If $X \geq 0$:

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

Probability of being large is controlled by mean.

49.2 Chebyshev's inequality

Corollary 49.3. Let X be a random variable with finite mean and variance. Then for any $a > 0$,

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

Remark. Chebyshev's inequality gives an upper bound of the tail probability $P(|X - \mathbb{E}[X]| \geq a)$.

- The error (how X_n far from its mean) is capped by its variance

50 Law of Large Numbers

Theorem 50.1 (Strong Law of Large Numbers)**.** Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ and suppose $E[|X_1|] < \infty$. Then there exists a set $A \subset \Omega$ such that $P(A) = 1$ and, for every $\omega \in A$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) = \mu.$$

Theorem 50.2 (Weak Law of Large Numbers)**.** Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with finite mean μ and variance σ^2. Then, for every $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Remark. Comparisons of strong and weak LLN:

1. Strong LLN: whole sequence, every single realized sequence converges.
$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu$$
2. Weak LLN: Converges in probability
$$\bar{X}_n \xrightarrow{P} \mu$$

a.s. : almost surely (stronger)

50.1 Sums and Averages

1. Setup and notation
Let $X_1, X_2, \dots$ be independent and identically distributed (i.i.d.) random variables.
Mean: $\mathbb{E}[X_i] = \mu$
Variance: $\text{var}(X_i) = \sigma^2$

- Sum: $S_n = X_1 + \cdots + X_n$
- Sample mean: $\bar{X}_n = S_n/n$

50.1.1 Mean and variance of sums and averages

Sums: $\mathbb{E}[S_n] = n\mu$, $\text{var}(S_n) = n\sigma^2$, $\sigma(S_n) = \sigma\sqrt{n}$
Sample mean: $\mathbb{E}[\bar{X}_n] = \mu$, $\text{var}(\bar{X}_n) = \sigma^2/n$, $\sigma(\bar{X}_n) = \sigma/\sqrt{n}$

Interpretation
The average is unbiased (correct on average). Its variability shrinks like $1/n$ (raw engine behind Law of Large Number).

50.2 Monte Carlo estimation land

In this example,

$$\theta = E[f(X)].$$

- X is a random variable from which we can simulate (draw samples).
- $f(X)$ is a transformation of X (e.g., a square, exponential, indicator, or other function).
- θ is the real mean of the transformed values $f(X)$.

Think of θ as the unknown quantity we want to estimate but cannot compute directly.

Theorem 50.3 (Monte Carlo)**.** X follows some distribution that we can simulate. Suppose we aim to estimate

$$\theta = E[f(X)]$$

We generate i.i.d. samples $X_1, \ldots, X_n$ and form the estimator

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The Strong Law of Large Numbers guarantees that

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

51 Symbols

Symbol	Meaning
X	The random rule (random variable as a function)
X_i	The ith random variable generated by the rule
$X_i(\omega)$	The actual observed value (data realization)

Probability land	Data land
Random variable (X)	Observed values ($X(\omega)$)
Abstract function	Concrete numbers
We talk about **distribution**	We talk about **data**
All branches	One chosen path

Remark. A realized sample:

- Before simulation, X is a function
- After simulation, each X_i becomes a real number

52 Power Law

Definition 52.1. A power-law distribution has probabilities that decay like a power of x,

$$f(x) = (\rho - 1)\, x^{-\rho}, \qquad x \geq 1, \qquad \rho > 1.$$

- Large values of X are rare, but not too rare.
- The tail decays slowly, compared to exponential or normal distributions.

52.1 The Tail behaviour

For large x,

$$f(x) \propto x^{-\rho}.$$

This creates a **heavy tail**: extreme values occur more often than intuition expects.

Part XII
Useful theorems

53 Binomial theorem

Theorem 53.1 (Binomial theorem)**.** For any interger $n0$,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k,$$

The binomial theorem naturally models an experiment with two possible outcomes each trial, repeated n times.

- x is the probability of success outcome (p)
- y is the probability of failure outcome ($1-p$)
- $\binom{n}{k}$ is the number of sequences with k successes and $n-k$ failures.

Part XIII
Markov Chain

54 A Stochastic process

Definition 54.1 (Stochastic Process)**.** A stochastic process is just a **sequence of random variables** that evolves over time: $X_0, X_1, \ldots$

55 Markov Chain

Remark. Markov chain is a type of Stochastic process. It assumes: only today matters. Once you know where you are now, the past is irrelevant.

55.1 Markov Property

Definition 55.1. The stochastic process $X_t = X_0, X_1, X_2, \ldots$ is a *discrete-time Markov chain* if it satisfies the **Markov property**:

$$\Pr(X_{n+1} = s \mid X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \Pr(X_{n+1} = s \mid X_n = x_n)$$

for all $x_0, x_1, \ldots, s \in S$ and for all $n \geq 0$.

55.2 Transition probability

Definition 55.2 (Transition probabilities)**.** Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with state space S, where S has size N (possibly infinite). The *transition probabilities* of the Markov chain are

$$p_{ij} = \Pr(X_{t+1} = j \mid X_t = i), \qquad i, j \in S,\ t = 0, 1, 2, \ldots$$

56 Transition Matrix

56.1 One step

Definition 56.1 (Transition Matrix (One step))**.** The transition matrix of the Markov chain is $P = (p_{ij})$.

$$\text{Transition Matrix } P = (p_{ij}) = \begin{pmatrix} \text{list all states } X_{t+1} & & & \\ p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \text{list all states } X_t$$

- Insert transition probabilities p_{ij} into the matrix.
- Each **row** represents state NOW, or from X_t. It is a full distribution of next-step probabilities given that you'are in state i now.

– The rows of P should each sum to 1:

$$\sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} \Pr(X_{t+1} = j \mid X_t = i) = \sum_{j=1}^{N} P_{\{X_t=i\}}(X_{t+1} = j) = 1.$$

– $p_{ij} \geq 0$, since they are probabilities.
– $j = 1, 2, \ldots, N$ are just labels for states

- Each **column** corresponds to the NEXT state X_{t+1} and do not generally sum to 1.

57 n steps

Definition 57.1 (Transition matrix (m steps))**.** Given a Markov chain X_n with transition matrix p, then the m-step transition probability of moving from state i to state j in exactly m steps is:

$$p^{(m)}(i,j) = \Pr\big(X_{n+m} = j \,\big|\, X_n = i\big),$$

- Note that steps $= (n+m) - n = m$

Theorem 57.2 (The Chapman–Kolmogorov equation)**.** The m-th step transition probability is:

$$p^{(m+n)}(i,j) = \sum_k p^{(m)}(i,k)\, p^{(n)}(k,j),$$

- In matrix form: $P^{(m)} = P^m$
- Sum over N possible intermediate states
- Number of steps = m

Remark. With a process depends on more than 1 past states, then treat all past states as a "block" of all history information to make them again 1st-order Markov chain.

58 Stationary Distribution

Definition 58.1. A Markov chain consists of:

- A finite set of states $\{1, 2, \ldots, N\}$.
- A transition matrix $P = (p_{ij})$ where $p_{ij} = P(X_{t+1} = j \mid X_t = i)$.
- A sequence of random variables $X_0, X_1, X_2, \ldots$, each taking values in $\{1, \ldots, N\}$.
- At each time t, the distribution of X_t is represented by a probability vector

$$\pi^{(t)} = (\pi_1^{(t)}, \pi_2^{(t)}, \ldots, \pi_N^{(t)}) \text{ where } \pi_i^{(t)} = P(X_t = i)$$

58.1 The distribution of X_0

Formula 58.2. We describe the probability of state at starting time using a probability vector:

$$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} = \begin{pmatrix} P(X_0 = 1) \\ P(X_0 = 2) \\ \vdots \\ P(X_0 = N) \end{pmatrix}.$$

- The row-vector version is: $\pi^T = (\pi_1, \pi_2, \ldots, \pi_N)$.
- Notation $X_0 \sim \pi^T$: the distribution of X_0 is given by the row vector π^T.

58.2 Distribution of X_1

Formula 58.3. At time $t = 1$, X_1's state depends on X_0's:

$$X_0 : P(X_1 = j) = \sum_{i=1}^{N} P(X_1 = j \mid X_0 = i)\, P(X_0 = i).$$

Interpretation: to be in state j at time 1, you must come from some previous state i, with probability p_{ij}, and the previous state i had probability π_i.

$$P(X_1 = j) = \sum_{i=1}^{N} \pi_i\, p_{ij}.$$

58.3 2 statesp

59 State properties

We observe a system with full past is $\mathbf{X} = [X_n, X_{n-1}, \ldots, X_0]$ where

- at times $n = 0, 1, 2, ...$
- at states labeled by numbers: $i = 0, 1, 2, ...$
- At each time n, the system is in some random state X_n. State is time-dependent and random.

The process $\{X_n\}$ is called a Markov chain if:

$$\Pr(X_{n+1} = j \,|\, X_n = i,\, X_{n-1}) \;=\; \Pr(X_{n+1} = j \,|\, X_n = i) \;=\; P_{ij}.$$

Definition 59.1. Let f_i be the probability that starting at state i, Markov chain ever reenters state i.

$$f_i \;:=\; \Pr\left(\bigcup_{n=1}^{\infty} \{X_n = i\} \;\middle|\; X_0 = i \right) \;=\; \Pr\left(\bigcup_{n=m+1}^{\infty} \{X_n = i\} \;\middle|\; X_m = i \right).$$

59.1 Recurrent State

Definition 59.2 (Recurrent state)**.** State i is recurrent if $p_{ii} = 1$, returning to state i is guaranteed.

Remark. A recurrent state is a long-run property, not about how fast but about whether we come back state i.

59.2 Transient State

Definition 59.3 (Transient state)**.** State i is transient if $p_{ii} < 1$, returning to state i is not guaranteed. The probability of never combing back to i is $1 - pii$.

Accessbile State

Definition 59.4. A state j is said to be *accessible* from a state i if

$$P_{ij}^{(n)} > 0 \quad \text{for some } n \geq 0.$$

Starting from state i, there is a positive probability that the Markov chain reaches j in some number of steps.

Remark. Every state is accessible from itself: $p_{ii}^{(0)} = 1$.

59.3 Absorbing State

Definition 59.5 (Absorbing State)**.** A state is called absorbing of $P_a(X_1 = a) = 1$.

60 Chain properties

60.1 Communicate

Definition 60.1 (Communicate)**.** States i and j are said to *communicate* (written $i \leftrightarrow j$) if

1. i is accessible from j, that is, $p_{ij}^n > 0$ for some $n \geq 0$
2. j is accessible from i, that is, $p_{ji}^m > 0$ for some $m \geq 0$.

Property 61. Communication is an equivalence relation because:

- it is **reflexive**: every state can reach itself, so $i \leftrightarrow i$;
- it is **symmetric**: if i can reach j and j can reach i, then $i \leftrightarrow j$ implies $j \leftrightarrow i$;
- it is **transitive**: if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$.

61.0.1 Irreducible

Definition 61.1. A set of states C is *irreducible* if every state can reach every other with positive probability. For all $i, j \in C$,

$$\rho_{ij} > 0,$$

62 Periodicity

62.1 Period

Definition 62.1 (Period of a state)**.** The period k of a state i is defined as

$$k \;=\; \gcd\{\, n \geq 1 : \Pr(X_n = i \mid X_0 = i) > 0 \,\}.$$

Periodicity is **state-specific**.

Remark. The period is based on the GCD of the return times, not necessarily one of the return times itself. e.g. gcd(t=6,8,10,12)=2 where 6,8,10,12,... is a list of possible return times (multiple of 2), but 2 is period and not in the list.

62.2 Aperiodic state

Definition 62.2. If the period of a state is k = 1, then the state is said to be aperiodic.

Remark. Aperiodic does NOT mean: You can return at every time (1,2,3,4,...), you return frequently, or state is absorbing. There is no strict timing rules, you can return state i at time t=4,5,.. or t=6,9,20,...

63 Long-term behaviour of Markov chains

Each state usually has its own long-run probability, and they are not equal.

Definition 63.1. Suppose there exists a distribution $p^* = (p_i)^*$ on state space S such that in long time, the prob. of reaching state j at time n eventually approaches a fixed number.

$$P(X_n = j) \longrightarrow p_j^*$$

Then p^* is an equilibrium distribution.

Theorem 63.2 (Limit theorem)**.** If a Markov chain is irreducible + aperiodic, where μ_j is the mean recurrence time of state j, then,

$$P(X_n = j) \longrightarrow \frac{1}{\mu_j},$$

1. **Case 1 - Every state has a finite expected return time:**

$$\pi_j = \frac{1}{\mu_j}$$

   The distribution $\pi = (\pi_j)$ is unique and equilibrium.

2. **Case 2 - Null recurrent or transient:** $\mu_j = \infty$, $\frac{1}{\mu_j} = 0$, for all states j

$$P(X_n = j) \longrightarrow 0$$

   No equilibrium distribution in this case.

Remark. The long–run probability of being in state j is inversely related to how long it takes to return to j:

- short return time ⇒ large probability
- long return time ⇒ small probability.

The limit theorem $p_{ij}^{(n)} \longrightarrow \pi_j$ for all i and j tells us that the n-step transition matrix has the limiting value where each row is identical.

$$\lim_{n\to\infty} P^{(n)} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \\ \pi_1 & \pi_2 & \cdots & \pi_N \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_N \end{pmatrix},$$

64 Expected holding time

Assume the chain is discrete-time.

From state i:

- With probability $P(i \to i)$, you stay in i.
- With probability $1 - P(i \to i)$, you leave i.

At each step:

- you either *fail* (stay), or
- you *succeed* (leave),

Formula 64.1.

$$T_i \sim \text{Geometric}\big(1 - P(i \to i)\big).$$

65 Common theorems

Theorem 65.1. If a Markov chain is irreducible, then there exists a unique stationary distribution $\pi(x)$, and every state has positive probability, i.e.,

$$\pi(y) > 0 \qquad \text{for all } y.$$

Theorem 65.2. For a two–state irreducible Markov chain, the stationary distribution can be computed explicitly. We have

$$p_n - \frac{b}{a+b} \;=\; \left(p_0 - \frac{b}{a+b}\right) \left|\, 1 - a - b \,\right|^n. \tag{6.9}$$

If $0 < a + b < 2$, this implies that the convergence is *exponentially fast*.

$$p_n \;\longrightarrow\; \frac{b}{a+b}$$

1. LHS = Error at time n, the non-equilibrium of the system at time n.
2. RHS: initial error x a shrinking factor = rate of shrinking.

Moreover, the convergence rate is known exactly:

$$\text{error after } n \text{ steps} \;=\; \left|\, 1 - a - b \,\right|^n.$$

- the stationary distribution exists,
- it is unique,
- and the chain converges to it *exponentially fast*.

Remark. In many models (weather, genetics, queues), convergence is so fast that a few steps is enough to approximate "long-run" behavior.

Theorem 65.3 (Exit times)**.** Let a set of state A be absorbing states, we define the exit time as number of steps until the chain first enters A:

$$V_A = \inf\{n \geq 0 : \, X_n \in A\}.$$

and the expected time to hit A start from state x not in A:

$$g(x) = \mathbb{E}_x[V_A]$$

Assume $S \setminus A$ is finite and $P_x(V_A < \infty) > 0$ for all $x \in S \setminus A$. Then, for every $x \in S \setminus A$,

$$\begin{aligned} h(x) &= b + \sum_{y \in S} r(x,y)\, h(y) \\ &= 1 + r\, g(x) \end{aligned}$$

- Take one step immediately at current state x
- We'll reach some state y with probability $p(x,y)$
- From that state y, it will take on average $g(y)$ more steps to reach A

Remark. Depend on what quantity is measured:

$$h = (I - r)^{-1} b, \tag{65.1}$$

$$g = (I - r)^{-1} 1. \tag{65.2}$$

1. $h(x)$ is a *probability*:

$$h(x) = \Pr(\text{eventually absorbed in a specified absorbing state} \mid X_0 = x), \quad 0 \le h(x) \le 1.$$

   (a) Recursion:

$$\begin{aligned} h(x) &= b + \sum_{y \in S} r(x,y)\, h(y) \\ &= b + rh \end{aligned}$$

   (b) b is (jumping) probability of being in absorbing state from current state
   (c) $h(x)$ measures the chance of eventual success. If there are 2 absorbing states, $h(G) = 1, h(D) = 0$.

2. $g(x)$ is an *expected time periods*:

$$g(x) = \mathbb{E}[\text{steps until absorption} \mid X_0 = x], \quad g(x) \in [0, \infty].$$

   (a) Recursion:

$$\begin{aligned} g(x) &= 1 + \sum_{y \in S} r(x,y)\, g(y) \\ &= 1 + rg \end{aligned}$$

   (b) 1 is the time spent in current state
   (c) $g(x)$ measures how long absorption takes. If there are 2 absorbing states, $g(G) = g(D) = 0$.

Theorem 65.4. Some notations:

1. $p^n(i,j)$: the probability of being in transient state j at time n steps, including absorbing states
2. $r^n(i,j)$: the probability of being in transient state j at time n, without including absorbing states.

Remark (Path)**.** Matrix $(I - r)^n$ ALWAYS measures expected time periods spent in transient state j. It takes the average value of total time periods spent in state j from all possible paths.

$$(I - r)^{-1} = I + r + r^2 + r^3 + \cdots$$

1. $g = (I - r)^{-1}\mathbf{1}$: Total expected number of time periods spent in state j before absorption.
   - Each entry $(I - r)^{-1}(i,j)$ is the expected time periods spent in state j, starting from state i.
2. $h = (I - r)^{-1} b$: the probability of being in an absorbing state of interest, from state i.

Lemma 65.5. If p is irreducible then all states have the same period.

Lemma 65.6. If p is irreducible and p(x, x) ¿ 0 then x has period 1, and hence by the previous lemma all states have period 1.

66 Number of visits

Formula 66.1. Define

$$N(y) = \sum_{n=1}^{\infty} \mathbf{1}(X_n = y),$$

the total number of visits to state y after time 0.

- $N(y) = \infty$: the chain returns to y infinitely often (keeps coming back).
- $N(y) < \infty$: the chain visits y only finitely many times (eventually leaves forever).

67 Models

67.1 Random walk

Remark. The walker moves one unit at a time, like taking steps.

- Walk = sequence of steps
- Random = direction (left/right) chosen driven by random event

68 Examples

68.1 Ehrenfest Chain

- 2 urns
- Total N balls
- At each step, we pick one ball uniformly at random and move it to the other urn. Let X_n be the number of balls in the left urn at time n
- Suppose there are i balls in the left urn currently.

The transition probabilities are

1. i increases to $i+1$:
$$p(i, i+1) = \frac{N-i}{N}$$
2. i decreases to $i-1$:
$$p(i, i-1) = \frac{i}{N}, \qquad 0 \le i \le N,$$
3. Otherwise:
$$p(i, j) = 0 \quad \text{otherwise.}$$

68.2 Wright-Fisher Model

$$p(i, j) = \Pr(X_{n+1} = j \mid X_n = i) = \binom{N}{j} \rho_i^j \left(1 - \rho_i\right)^{N-j}. \tag{68.1}$$

- N genes with 2 types of alleles: a and A
- Next generation is sampled with replacement fron current generation = independent draws

$$X_n = i \quad \text{(i A-alleles at time } n\text{)}.$$

$$\rho_i = \frac{i}{N}(1-u) + \frac{N-i}{N}v.$$

Then the next state is distributed as

$$X_{n+1} \sim \text{Binomial}(N, \rho_i).$$

Therefore the one-step transition probability is

$$p(i,j) = \Pr(X_{n+1} = j \mid X_n = i) = \binom{N}{j} \rho_i^{\,j} (1-\rho_i)^{\,N-j}.$$

Note that the transition probability depends only on i and not on any earlier values (Marko property).

$$X_{n-1},\ X_{n-2},\ \ldots$$

Part XIV
Moment Generating Function

69 Expectation of a Function $E[r(X)]$

Definition 69.1. The expectation of a function $r(x)$ of a random variable X, denoted $E[r(X)]$, is the weighted average of the possible values of $r(x)$, where the weights are the probabilities associated with x.

69.1 Discrete Case

For a **discrete** random variable X with a pmf $P(X = x)$:

$$E[r(X)] = \sum_x r(x)\, P(X = x)$$

69.2 Continuous Case

For a **continuous** random variable X with a pdf $f(x)$:

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

70 Moments

Definition 70.1. Let funtion $r(X) = X^k$. For a random variable X and integer $k = 1, 2, \ldots$, the k-th moment is

$$m_k = E[X^k]$$

70.1 Uniqueness of Moments

Theorem 70.2. Let X and Y be two random variables taking values in a finite interval; that is, for some constant $M > 0$,

$$\mathbb{P}(-M < X < M) = 1 \qquad \text{and} \qquad \mathbb{P}(-M < Y < M) = 1.$$

If all moments are finite and

$$\forall\, k \geq 1, \quad \mathbb{E}[X^k] = \mathbb{E}[Y^k],$$

then X and Y have the same distribution.

Same distribution: same CDF, same density, same probability law, everything.

71 2nd Moment - Variance

71.1 Average Absolute Deviation

1. Distance from the mean: $|X - \mu|$
2. Average: $E[|X - \mu|]$

71.2 Standard Deviation

Instead of using the absolute distance, statistics uses **root-mean-square deviation**:

$$\sqrt{E[(X-\mu)^2]}$$

- Squaring emphasizes larger deviations.
- Squaring avoids the non-differentiability of $|x|$ at 0.

Remark. SD is slight larger than Average Absolute Deviation as squaring punishes big deviations more.

Remark. A moment measures how far and in what way the probability mass is spread out around some reference point (often 0 or the mean).