Probability Theory

Labix

April 11, 2025

Abstract

Notes for the basics of Probability Theory.

Contents

1 Probability with Set Language **2**
- 1.1 Definition of Probability . . . 2
- 1.2 Multiplication Principle . . . 3
- 1.3 Conditional Probability . . . 3
- 1.4 Independence of Events . . . 5

2 Distributions **6**
- 2.1 Random Variables . . . 6
- 2.2 Probability Distribution . . . 6
- 2.3 Common Discrete Distribution . . . 7
- 2.4 Common Continuous Distributions . . . 8
- 2.5 Transformation of Random Variables . . . 10
- 2.6 Multivariate Random Variables . . . 10
- 2.7 Distribution of Sums . . . 11

3 Expectation and Variance **16**
- 3.1 Expectations . . . 16
- 3.2 Variance . . . 17
- 3.3 Covariance . . . 17
- 3.4 Moments . . . 18

4 Convergence of Random Variables **20**
- 4.1 Convergence . . . 20
- 4.2 Standardized Random Variables . . . 21

5 Stochastic Processes **22**
- 5.1 Markov Chains . . . 22
- 5.2 Communicating Classes . . . 23
- 5.3 Hitting Times . . . 24
- 5.4 Strong Markov Property . . . 24
- 5.5 Recurrence and Transcience . . . 25
- 5.6 Branching Process . . . 26
- 5.7 Invariant Distributions on a Markov Chain . . . 27

1 Foundations of Probability Theory

1.1 Definition of Probability

Definition 1.1.1: Probability Space

A probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of the following data:
- $\Omega \neq \emptyset$ is a set called the sample space.
- $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra called events.
- $P : \mathcal{F} \to [0, 1]$ is a set function.

such that the following are true:
- $P(\Omega) = 1$.
- If $\{A_n | n \in \mathbb{N}\} \subseteq \mathcal{F}$ are pairwise disjoint, then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$$

Proposition 1.1.2

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $A, B \in \mathcal{F}$ be events. Then the following are true.
- $P(\Omega \setminus A) = 1 - P(A)$
- $A \subset B \implies P(A) \leq P(B)$

Proof. Let $A \subset B \subset \Omega$ be events in Ω.
- A and $\Omega \setminus A$ are disjoint and $P(\Omega) = P(A) + P(\Omega \setminus A)$ and $P(\Omega \setminus A) = 1 - P(A)$
- We have that A and $B \setminus A$ are disjoint. Thus $P(B) = P(A) + P(B \setminus A)$. Since $P(B \setminus A) \geq 0$, we have $P(A) \leq P(B)$.

□

Definition 1.1.3: Uniform Probability Measure

Let Ω be a sample space. A probability measure P is uniform if fo all $a, b \in \Omega$,

$$P(\{a\}) = P(\{b\})$$

Theorem 1.1.4

Let Ω be a sample space and P a uniform probability measure of Ω. Then for all $A \subset \Omega$,

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

Proof. Suppose that A consists of $|A|$ distinct elements and the event space $|\Omega|$ contains $|\Omega|$ distinct elements. Since every singleton set is pairwise disjoint, we have $P(A) = |A|P(\{a\})$ for any $a \in A$. Similarly, we have $P(\Omega) = |\Omega|P(\{a\})$. Thus we have that $P(A) = \frac{|A|P(\Omega)}{|\Omega|}$ and $P(A) = \frac{|A|}{|\Omega|}$ □

Theorem 1.1.5: Principle of Inclusion Exclusion

Let $A, B \subset \Omega$ be a sample space and P the probability measure.

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

Proof. Note that

$$\begin{aligned} A \cup (B \setminus A) &= A \cup (B \cap A^c) \\ &= (A \cup B) \cap (A \cup A^c) \\ &= A \cup B \end{aligned}$$

Note also that $A \cap (B \setminus A) = \emptyset$. Thus $P(A \cup B) = P(A) + P(B \setminus A) = P(A) + P(B) - P(A \cap B)$ □

Theorem 1.1.6: Extended Principle of Inclusion Exclusion

Let $A_k \subset \Omega$ be a sample space and P the probability measure for all $k \leq n \in \mathbb{N}$. Then

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 \leq \cdots \leq n} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k})$$

1.2 Multiplication Principle

Theorem 1.2.1: The Multiplication Principle

Suppose that Experiment A has a outcomes and Experiment B has b outcomes. Then the performing both A and B results in ab possible outcomes.

Theorem 1.2.2: Sampling with replacement - Ordered

In the case of sampling k balls with replacement from an urn containing n balls, there are $|\Omega| = n^k$ possible outcomes when the order of the objects matters, where $\Omega = \{(s_1, \ldots, s_k) : s_i \in \{1, \ldots, n\} \forall i \in \{1, \ldots, k\}\}$.

Theorem 1.2.3: Sampling without replacement - Ordered

In the case of sampling k balls without replacement from an urn containing n balls, there are $|\Omega| = \frac{n!}{(n-k)!}$ possible outcomes when the order of the objects matters, where $\Omega = \{(s_1, \ldots, s_k) : s_i \in \{1, \ldots, n\} \forall i \in \{1, \ldots, k\}, i \neq j \implies s_i \neq s_j\}$.

Theorem 1.2.4: Sampling without replacement - Unordered

In the case of sampling k balls without replacement from an urn containing n balls, there are $|\Omega| = \binom{n}{k}$ possible outcomes when the order of the objects does not matter, where $\Omega = \{\omega \subset \{1, \ldots, n\} : |\omega| = k\}$.

Theorem 1.2.5: Sampling with replacement - Unordered

In the case of sampling k balls with replacement from an urn containing n balls, there are $|\Omega| = \binom{n+k-1}{k}$ possible outcomes when the order of the objects does not matter, where $\Omega = \{\omega \subset \{1, \ldots, n\} : \omega \text{ is a } k \text{ element multiset with elements from } \{1, \ldots, n\}\}$.

1.3 Conditional Probability

Definition 1.3.1: Conditional Probability

Consider a probability space (Ω, P). Let $A, B \subset \Omega$ with $P(B) > 0$. Then the conditional

probability of A given B, denoted by $P(A|B)$ is defined as

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

Theorem 1.3.2: Multiplication Rule

Let $n \in \mathbb{N}$. Then for any events $A_1, \ldots, A_n$ such that $P(A_2 \cap \cdots \cap A_n) > 0$, we have

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

Proof. From the right hand side, we have

$$\begin{aligned} &P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}) \\ &= P(A_1)\frac{P(A_2 \cap A_1)}{P(A_1)}\frac{P(A_3 \cap A_2 \cap A_1)}{P(A_2 \cap A_1)} \cdots \frac{P(A_n \cap \cdots \cap A_1)}{P(A_1 \cap \cdots \cap A_{n-1})} \\ &= P(A_1 \cap \cdots \cap A_n) \end{aligned}$$

□

Theorem 1.3.3: Bayes' Rule

Let (Ω, P) be a probability measure. Let $A, B \subset \Omega$ with $P(A), P(B) > 0$. Then

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

Proof. We have that $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B) = P(B|A)P(A)$. □

Theorem 1.3.4: Law of Total Probability

Let (Ω, P) be a probability measure. Let $A_1, \ldots, A_n$ be a partition of Ω with $P(A_i) > 0$ for all $i = 1, \ldots, n$. Then for any $B \subset \Omega$,

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B|A_k)$$

Proof. Note that since $A_1, \ldots, A_n$ is a partition, $B \cap A_1, \ldots, B \cap A_n$ is also a parition.

$$\begin{aligned} \sum_{k=1}^{n} P(A_k)P(B|A_k) &= \sum_{k=1}^{n} P(B \cap A_k) \\ &= P\left(\bigcup_{k=1}^{n} B \cap A_k\right) \\ &= P(B \cap \Omega) \\ &= P(B) \end{aligned}$$

□

Theorem 1.3.5: General Bayes' Rule

Let (Ω, P) be a probability measure. Let $A_1, \ldots, A_n$ be a partition of Ω with $P(A_i) > 0$ for all

$i = 1, \ldots, n$. Then for any $B \subset \Omega$ with $P(B) > 0$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_i)P(A_i)}$$

Proof. Apply Bayes' rule and apply the mulitplication rule. □

1.4 Independence of Events

Definition 1.4.1: Independent Events

Two events A, B are said to be independent if

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

Proposition 1.4.2

If A, B are independent, then A^c, B, A, B^c and A^c, B^c are independent.

Proof. We only proof the first and third item.

- Without loss of generality we prove the first and reader mirrors the second.

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) \\ &= P(B)(1 - P(A)) \\ &= P(B)P(A^c) \end{aligned}$$

- Note that $P(A \cap B) = P(A)P(B)$

$$\begin{aligned} P(A^c \cap B^c) &= 1 - P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A)P(B) \\ &= (1 - P(A))(1 - P(B)) \\ &= P(A^c)P(B^c) \end{aligned}$$

□

2 Distributions

2.1 Random Variables

Definition 2.1.1: Random Variable

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(E, \mathcal{E})$ be a measurable space. An $(E, \mathcal{E})$ valued random variable is an $\mathcal{F}$-measurable function $X : \Omega \to E$.

Recall that X is an $\mathcal{F}$-measurable function if $X^{-1}(B) \in \mathcal{F}$ for $B \in \mathcal{E}$.

Definition 2.1.2: Probability Distribution

Let $(\Omega, E, \mathbb{P})$ be a probability space. Let $(E, \mathcal{E})$ be a measurable space. Let $X : \Omega \to E$ be a measurable function. Define the probability distribution of X to be the function

$$P(X \in A) = P(X^{-1}(A))$$

for $A \in \mathcal{E}$.

Definition 2.1.3: Discrete Random Variable

A discrete random variable on the probability space $(\Omega, E, \mathbb{P})$ is a random variable such that $\text{im}(X) = \{X(\omega) : \omega \in \Omega\}$ is a countable subset of $\mathbb{R}$.

Definition 2.1.4: Continuous Random Variable

A continuous random variable on a probability space $(\Omega, E, \mathbb{P})$ is a random variable if there exists a non-negative function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{P}_X(A) = \mathbb{P}_X(X^{-1}(A)) = \int_A f_X(r)\, dr$$

2.2 Probability Distribution

Definition 2.2.1: Probability Mass Function

The probability mass function of the discrete random variable X is defined as the function $p_X : \mathbb{R} \to [0, 1]$ given by

$$p_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(X = x)$$

Lemma 2.2.2

Let X be a random variable and p_X the probability mass function. Then

$$\sum_{x \in \text{im}(X)} p_X(x) = 1$$

Proof. It is on a probability space. □

Definition 2.2.3: Cumulative Distribution Function

Suppose that X is a random variable on a probability space $(\Omega, E, \mathbb{P})$. Then the cumulative distribution function of X is defined as the mapping $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X^{-1}(-\infty, x]) = \mathbb{P}_X(X \leq x)$$

Lemma 2.2.4

If X is a discrete random variable, then the cummulative distribution function of X is

$$F_X(x) = \sum_{k \leq x} p_X(k)$$

If X is a continuous random variable, then the cummulative distribution function of X is

$$F_X(t) = \int_{-\infty}^{t} f(x)\,dx$$

Theorem 2.2.5

A comparison of the two probability functions with discrete random variable and continuous random variable.

	Discrete	Continuous
Probability Function	$p_X(x) \geq 0$ for all $x \in \mathbb{R}$	$f_X(x) \geq 0$ for all $x \in \mathbb{R}$
Cumulative Probability	$\sum_{x \in \text{im}(X)} p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)dx = 1$
Cumulative Distribution Function	$F_X(x) = \sum_{x \in \text{im}(X): u \leq x} p_X(u)$	$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$

Theorem 2.2.6

Suppose that X is a random variable on a probability space $(\Omega, E, \mathbb{P})$ with cumulative distribution function F_X.

- F_X is monotonically non-decreasing. $x \leq y \implies F_X(x) \leq F_X(y)$
- F_x is right continuous. If (x_n) is a sequence such that $x_1 \geq \cdots \geq x_n \geq x_{n+1} \geq \cdots \geq x$ and $(x_n) \to x$, then $F_X(x_n) \to F_X(x)$
- $F_X(-\infty) = 0$ and $F_X(\infty) = 1$

Theorem 2.2.7

Suppose that X is a random variable on a probability space $(\Omega, E, \mathbb{P})$ with cumulative distribution function F_X. If $a < b$, then $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$

Theorem 2.2.8

For a continuous random variable X iwth density f_X, we have

- $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$
- $\mathbb{P}(a \leq x \leq b) = \int_a^b f_X(u)du$ for all $a, b \in \mathbb{R}$ with $a \leq b$

2.3 Common Discrete Distribution

Definition 2.3.1: Bernoulli Distribution

A discrete random variable X is said to have Bernoulli Distribution with parameter $p \in (0, 1)$ if $\text{im}(X) = \{0, 1\}$ and $p_X(1) = p$ and $p_X(0) = 1 - p$.

Definition 2.3.2: Binomial Distribution

A discrete random variable X is said to have Binomial Distribution with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$ if $\text{im}(X) = \{0, 1, \ldots, n\}$ and

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

Theorem 2.3.3: Bernoulli's Weak Law of Large Numbers

Let $p \in (0,1)$. For all $n \in \mathbb{N}$, let X_n have the binomial distribution with parameters n, p. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{X_n}{n} - p\right| > \epsilon\right) = 0$$

In other words, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(np - n\epsilon \leq X_n \leq np + n\epsilon) = 1$$

Definition 2.3.4: Poisson Distribution

A discrete random variable X is said to have Poisson Distribution with parameter $\lambda > 0$ if $\text{im}(X) = \mathbb{N}_0$ and

$$p_X(x) = \frac{\lambda^x}{x!}e^{-x}$$

Theorem 2.3.5: Poisson Approximation

Let (p_n) be a sequence with $p_n \in [0,1]$ for all n and $\lim_{n\to\infty} np_n = \lambda > 0$. Denote $X \approx$Poisson(λ) and $X_n \approx$Bin(n, p_n). Then for every $x \in \mathbb{N}$,

$$\lim_{n\to\infty} p_{X_n}(x) = p_X(x)$$

Proof.

$$\begin{aligned}
\lim_{n\to\infty} p_{X_n}(x) &= \lim_{n\to\infty} \binom{n}{x}(p_n)^x(1-p_n)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!}(p_n)^x(1-p_n)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \lim_{n\to\infty} \frac{n!}{(n-x)!}\left(\frac{1}{n^k}\right)\left(\frac{\lambda^x}{x!}\right)\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}e^{-x}
\end{aligned}$$

□

Definition 2.3.6: Geometric Distribution

A discrete random variable X is said to have Geometric Distribution with parameter $p \in (0,1)$ if $\text{im}(X) = \mathbb{N}_0$ and

$$p_X(x) = p(1-p)^{x-1}$$

2.4 Common Continuous Distributions

Definition 2.4.1: Uniform Distribution

A conitnuous random variable X is said to have Uniform Distribution on the interval (a,b) if

its density function is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

and its cumulative function given by

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

Definition 2.4.2: Exponential Distribution

A conitnuous random variable X is said to have Exponential Distribution with parameter $\lambda > 0$ if its density function is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and its cumulative function given by

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

Definition 2.4.3: Gamma Distribution

A conitnuous random variable X is said to have Gamma Distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ if its density function is given by

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Definition 2.4.4: Normal Distribution

A conitnuous random variable X is said to have Normal Distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if its density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for all $x \in \mathbb{R}$.

Theorem 2.4.5: De Moivre-Laplace

Let X_n denote the binomial distrbution with parameters n, p for all $n \in \mathbb{N}$. For all $-\infty \leq z_1 < z_2 \leq \infty$,

$$\lim_{n\to\infty} P\left(np + z_1\sqrt{np(1-p)} \leq X_n \leq np + z_2\sqrt{np(1-p)}\right) = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

2.5 Transformation of Random Variables

Theorem 2.5.1

Let X be a discrete random variable on (Ω, P) and let $g : \mathbb{R} \to \mathbb{R}$ denote a deterministic function. Then $Y = g(X)$ is a discrete random variable with probability mass function given by

$$p_Y(y) = \sum_{x \in \text{im}(X) : g(x) = y} P(X = x)$$

for all $y \in \text{im}(Y)$ and 0 otherwise.

Theorem 2.5.2

Suppose that X is a continuous random variable with density f_X and $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing or decreasing and differentiable with inverse function denoted g^{-1}, then $Y = g(X)$ has density

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

for all $y \in \mathbb{R}$

2.6 Multivariate Random Variables

Definition 2.6.1: Joint Probability Mass Function

Let X, Y be discrete random variables. The joint probability mass function of X and Y is the function

$$p_{X,Y}(x, y) = P(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}) = P((X, Y) = (x, y))$$

for all $(x, y) \in \mathbb{R}^2$

Theorem 2.6.2

Let $p_{X,Y}$ be the joint probability mass function of two random variables X, Y.

- $p_X(x) = \sum_y p_{X,Y}(x, y)$
- $p_Y(y) = \sum_x p_{X,Y}(x, y)$

Definition 2.6.3: Joint Cumulative Distribution Function

Let X, Y be random variables. The joint cumulative distribution function of X and Y is the function

$$F_{X,Y}(x, y) = P(\{\omega \in \Omega : X(\omega) \leq x, Y(\omega) \leq y\}) = P(X \leq x, Y \leq y)$$

for all $(x, y) \in \mathbb{R}^2$

Theorem 2.6.4

Let $F_{X,Y}$ be the joint cumulative distribution function of two random variables X, Y.

- $\lim_{x,y \to -\infty} F_{X,Y}(x, y) = 0$
- $\lim_{x,y \to \infty} F_{X,Y}(x, y) = 1$
- $x \leq x'$ and $y \leq y'$ implies $F_{X,Y}(x, y) \leq F_{X,Y}(x', y')$
- $F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y)$
- $F_Y(x) = \lim_{x \to \infty} F_{X,Y}(x, y)$

Definition 2.6.5: Jointly Continuous

Let X, Y be random variables. X and Y are jointly continuous if

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du$$

for a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying

- $f_{X,Y}(u,v) \geq 0$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv\,du = 1$

We call $f_{X,Y}$ the joint density function of (X,Y).

Theorem 2.6.6

Let $F_{X,Y}$ be the joint cumulative distribution function of two random variables X, Y.

- $f_{X,Y}(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) & \text{if the derivative exists at } (x,y) \\ 0 & \text{otherwise} \end{cases}$
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$
- $f_Y(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$

Definition 2.6.7

Two random variables X, Y defined in the same probability space are said to be independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

Theorem 2.6.8

Two random variables X, Y defined in the same probability space are said to be independent if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

Theorem 2.6.9

Two discrete random variables X, Y defined in the same probability space are said to be independent if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

Theorem 2.6.10

Two continuous random variables X, Y defined in the same probability space are said to be independent if

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

2.7 Distribution of Sums

Theorem 2.7.1: Sum of Discrete Random Variables

Let X, Y be discrete random variables with density function $p_{X,Y}$. Then $Z = X + Y$ has density function given by

$$p_Z(m) = \sum_{k\in\mathbb{Z}} p_{X,Y}(k, m-k)$$

In particular, if X, Y are independent, then

$$p_Z(m) = \sum_{k\in\mathbb{Z}} p_X(k)p_Y(m-k)$$

Proposition 2.7.2

Let $X \approx \text{Poi}(\lambda)$ and $Y \approx \text{Poi}(\mu)$ be independent. $X + Y \approx \text{Poi}(\lambda + \mu)$.

Proof.

$$\begin{aligned}
p_{X+Y}(m) &= \sum_{k\in\mathbb{Z}} \frac{\lambda^k}{k!}e^{-k}\frac{\mu^{m-k}}{(m-k)!}e^{k-m} \\
&= \frac{1}{m!}e^{-m}\sum_{k=0}^{m} m!\frac{\lambda^k}{k!}\frac{\mu^{m-k}}{(m-k)!} \\
&= \frac{1}{m!}e^{-m}\sum_{k=0}^{m}\binom{m}{k}\lambda^k\mu^{m-k} \\
&= \frac{(\lambda+\mu)^m}{m!}e^{-m}
\end{aligned}$$

□

Proposition 2.7.3

Let $X_1, \ldots, X_n \approx \text{Bern}(p)$ be independent. $X_1 + \cdots + X_n \approx \text{Bin}(n, p)$.

Proof. We prove by induction. When $n = 2$,

$$\begin{aligned}
p_{X_1+X_2}(0) &= p_{X_1}(0)p_{X_2}(0) \\
&= 1 - 2p + p^2 \\
p_{X_1+X_2}(1) &= p_{X_1}(0)p_{X_2}(1) + p_{X_1}(1)p_{X_2}(0) \\
&= (1-p)(p) + p(1-p) \\
&= 2p(1-p) \\
p_{X_1+X_2}(2) &= p_{X_1}(0)p_{X_2}(2) + p_{X_1}(1)p_{X_2}(1) + p_{X_1}(2)p_{X_2}(0) \\
&= p^2 \\
p_{\text{Bin}(2,p)}(x) &= \binom{2}{x}p^x(1-p)^{n-x}
\end{aligned}$$

For $x \in \{0, 1, 2\}$, the two probability density functions match thus for the case $n = 2$, it is true. Now

suppose that $X_1 + \cdots + X_{n-1} \approx \text{Bin}(n-1, p)$. Let $Y = \text{Bin}(n-1, p) + X_n$. For $m \in \{0, \ldots, n\}$,

$$\begin{aligned}
p_Y(m) &= \sum_{k \in \mathbb{Z}} p_{\text{Bin}(n-1,p)}(k) p_{X_n}(m-k) \\
&= \sum_{k=0}^{m} p_{\text{Bin}(n-1,p)}(k) p_{X_n}(m-k) \\
&= \sum_{k=0}^{m} \binom{n-1}{k} p^k (1-p)^{n-1-k} p_{X_n}(m-k) \\
&= \sum_{k=m-1}^{m} \binom{n-1}{k} p^k (1-p)^{n-1-k} p_{X_n}(m-k) \\
&= \binom{n-1}{m-1} p^{m-1} (1-p)^{n-m} p_{X_n}(1) + \binom{n-1}{m} p^m (1-p)^{n-1-m} p_{X_n}(0) \\
&= \binom{n-1}{m-1} p^m (1-p)^{n-m} + \binom{n-1}{m} p^m (1-p)^{n-m} \\
&= \binom{n}{m} p^m (1-p)^{n-m}
\end{aligned}$$

Thus for the case $X_1 + \cdots + X_n$ it is true. □

Proposition 2.7.4

Let $X \approx \text{Bin}(m, p)$ and $Y \approx \text{Bin}(n, p)$ be independent. $X + Y \approx \text{Bin}(m+n, p)$.

Proof.

$$\begin{aligned}
p_{X+Y}(t) &= \sum_{k \in \mathbb{Z}} p_X(k) p_Y(t-k) \\
&= \sum_{k=0}^{t} \binom{m}{k} p^k (1-p)^{m-k} \binom{n}{t-k} p^{t-k} (1-p)^{n-t+k} \\
&= \sum_{k=0}^{t} \binom{m}{k} \binom{n}{t-k} p^t (1-p)^{m+n-t} \\
&= p^t (1-p)^{m+n-t} \sum_{k=0}^{t} \frac{m!}{k!(m-k)!} \frac{n!}{(t-k)!(n-t+k)!}
\end{aligned}$$

□

Theorem 2.7.5: Sum of Independent Continuous Random Variables

Let X, Y be independent continuous random variables with density function f_X and f_Y respectively. Then $Z = X + Y$ has density function given by

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

Proposition 2.7.6

Let $\lambda > 0$. Let $n \in \mathbb{N}$. Let $T_1, \ldots, T_n$ be independent random variables with exponential

distribution parameter λ. Then

$$Z = \sum_{k=1}^{n} T_k \approx \text{Gamma}(n, \lambda)$$

Proof. We prove by induction. When $n = 2$,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{T_1}(x) f_{T_2}(z-x)\,dx \\ &= \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)}\,dx \\ &= \lambda^2 e^{-\lambda z} \int_0^z dx \\ &= \lambda^2 z e^{-\lambda z} \end{aligned}$$

Thus the case $n = 2$ is true. Suppose that it is true for $n = k - 1$. Let $X \approx$Gamma$(n-1, \lambda)$.

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_{T_n}(z-x)\,dx \\ &= \int_0^z \frac{\lambda^{n-1}}{\Gamma(n-1)} x^{n-2} e^{-\lambda x} \lambda e^{-\lambda(z-x)}\,dx \\ &= \frac{\lambda^n}{\Gamma(n-1)} e^{-\lambda z} \int_0^z x^{n-2}\,dx \\ &= \frac{\lambda^n}{\Gamma(n-1)} e^{-\lambda z} \frac{1}{n-1} z^{n-1} \\ &= \frac{\lambda^n}{\Gamma(n)} z^{n-1} e^{-\lambda z} \end{aligned}$$

Thus we are done □

Proposition 2.7.7

Let $m, n \in \mathbb{N}$ and $\lambda > 0$. Let $X \approx$Gamma(m, λ) and $Y \approx$Gamma(n, λ) be independent. $X + Y \approx$Gamma$(m+n, \lambda)$.

Proof.

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx \\ &= \int_0^z \frac{\lambda^m}{\Gamma(m)} x^{m-1} e^{-\lambda x} \frac{\lambda^n}{\Gamma(n)} (z-x)^{n-1} e^{-\lambda(z-x)}\,dx \\ &= \frac{\lambda^{m+n}}{\Gamma(m)\Gamma(n)} e^{-\lambda z} \int_0^z x^{m-1}(z-x)^{n-1}\,dx \\ &= \frac{\lambda^{m+n}}{\Gamma(m)\Gamma(n)} e^{-\lambda z} \int_0^z x^{m-1} \sum_{k=0}^{n-1} \binom{n-1}{k} z^{n-1-k}(-x)^k\,dx \\ &= \frac{\lambda^{m+n}}{\Gamma(m)\Gamma(n)} e^{-\lambda z} \sum_{k=0}^{n-1} \binom{n-1}{k} z^{n-1-k}(-1)^k \int_0^z x^{m-1+k}\,dx \\ &= \frac{\lambda^{m+n}}{\Gamma(m)\Gamma(n)} z^{m+n-1} e^{-\lambda z} \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k \frac{1}{m+k} \end{aligned}$$

□

Theorem 2.7.8

Suppose that $T_1, T_2, \dots$ are independent random variables with exponential distribution parameter λ. Define for $t \geq 0$,

$$N_t = \begin{cases} 0 & \text{if } T_1 > t \\ 1 & \text{if } T_1 \leq t < T_1 + T_2 \\ 2 & \text{if } T_1 + T_2 \leq t < T_1 + T_2 + T_3 \\ \dots \end{cases}$$

Then, for any $t \geq 0$, we have that $N_t \approx \text{Poi}(\lambda t)$.

Definition 2.7.9: Poisson Process

The family of random variables $\{N_t : t \geq 0\}$ is said to be Poisson process of intensity λ if

- $N_0 = 0$
- for any $t_0, \dots, t_n$ with $0 = t_0 < t_1 < t_2 < \cdots < t_n$, the random variables N_{t_1}, $N_{t_2} - N_{t_1}$, $N_{t_3} - N_{t_2}$, $\dots$, $N_{t_n} - N_{t_{n-1}}$ are independent, and $N_{t_i} - N_{t_{i-1}} \approx \text{Poi}(\lambda(t_i - t_{i-1}))$

3 Expectation and Variance

3.1 Expectations

Definition 3.1.1: Expectation Random Variables

Let X be a random variable. The expectation of X is defined as

- $E(X) = \sum_{x \in \text{im}(X)} x\mathbb{P}_X(X = x)$ if X is discrete
- $E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$ if X is continuous

Proposition 3.1.2

Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$.

- If X is discrete then $E(g(X)) = \sum_{x \in \text{im}(X)} g(x)p_X(x)$ whenever it converges absolutely
- If X is continuous then $E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)\, dx$ whenever it converges absolutely

Proposition 3.1.3: Law of the Unconscious Staticians

Let $X_1, \ldots, X_n$ be random variables and $g : \mathbb{R}^n \to \mathbb{R}$ be a function.

- If $X_1, \ldots, X_n$ are discrete, then

$$E(g(X_1, \ldots, X_n)) = \sum_{\substack{x_1 \in \text{im}(X_1) \\ \vdots \\ x_n \in \text{im}(X)}} g(x_1, \ldots, x_n)p_X(x_1, \ldots, x_n)$$

- If $X_1, \ldots, X_n$ are continuous, then

$$E(g(X_1, \ldots, X_n)) = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n)f_X(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

Proposition 3.1.4

Assume that all the random variables that appear in the following statements have a well defined expectation.

- Suppose that X is a random variable such that there exists a constant c such that $P(X = c) = 1$. Then
$$E(X) = c$$
- If X, Y are random variables and $a, b \in \mathbb{R}$, then
$$E(aX + bY) = aE(X) + bE(Y)$$
- If X is a random variable such that $P(X \geq 0) = 1$, then $E(X) \geq 0$. If $P(X > 0) > 0$, then $E(X) > 0$. If X_1 and X_2 are random variables with $P(X_1 \geq X_2) = 1$, then $E(X_1) \geq E(X_2)$
- If X, Y are independent random variables, then
$$E(XY) = E(X)E(Y)$$

Theorem 3.1.5

Two random variables X, Y are independent if and only if

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

for any two functions $g, h : \mathbb{R} \to \mathbb{R}$ for which the expectation exists.

3.2 Variance

Definition 3.2.1: Variance

Let X be a random variable. The variance of X is defined as

$$\operatorname{Var}(X) = E((X - E(X))^2)$$

whenever the expectation exists. The standard deviation of X is defined as $\sigma_X = \sqrt{\operatorname{Var}(X)}$ whenever the variance exists

Theorem 3.2.2

Let X be a random variable whose variance is well defined.

- $\operatorname{Var}(X) \geq 0$
- $\operatorname{Var}(X) = 0$ if and only if $P(X = E(X)) = 1$
- $\operatorname{Var}(X) = E(X^2) - E(X)^2$

Proposition 3.2.3

Let X be a random variable whose variance is well defined.

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$$

for all $a, b \in \mathbb{R}$.

Theorem 3.2.4

Suppose that $X_1, \ldots, X_n$ are independent variables with finite variance. Then

$$\operatorname{Var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \operatorname{Var}(X_k)$$

Theorem 3.2.5

A list of expectations and variances of different distributions.

X	$E(X)$	$\operatorname{Var}(X)$
$\text{Ber}(p)$	p	$p(1-p)$
$\text{Bin}(n,p)$	np	$np(1-p)$
$\text{Poi}(\lambda)$	λ	λ
$\text{Geo}(p)$n	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\text{Uni}(a,b)$	$\frac{a+b}{2}$	$\frac{1}{12}(b-a)^2$
$\text{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$
$\text{Gamma}(\alpha,\beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
$\text{Norm}(\mu,\sigma^2)$	μ	σ^2

3.3 Covariance

Definition 3.3.1: Covariance

Let X, Y be two random variables. The covariance of X, Y is defined as

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

Proposition 3.3.2

Suppose that X, Y are random variables.
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
- If X, Y are independent, $\text{Cov}(X, Y) = 0$
- $\text{Cov}(aX + bY, Z) = a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z)$

Proposition 3.3.3: Variance of Sums

For random variables $X_1, \ldots, X_n$, we have

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

Theorem 3.3.4

Given two random variables X and Y, we have

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}$$

Theorem 3.3.5: Correlation Coefficient

The correlation coefficient between two random variables X and Y is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Proposition 3.3.6

Let X and Y be random variables. We have

$$-1 \leq \rho(X, Y) \leq 1$$

Moreover, for any $a, b, c, d \in \mathbb{R}$ with $a, c > 0$, we have

$$\rho(aX + b, cY + d) = \rho(X, Y)$$

Proposition 3.3.7

Let X, Y be random variables.
- $\rho(X, X) = 1$
- $\rho(X, -X) = -1$
- X, Y are uncorrelated if $\rho(X, Y) = 0$

3.4 Moments

Definition 3.4.1: kth Moment

Let X be a random variable. For $k \in \mathbb{N}$ we define the kth moment of X as $E[X^k]$ whenever the expectation exists.

Definition 3.4.2: Moment Generating Function

The moment-generating function of a random variable X is the function M_X defined as

$$M_X(t) = E[e^{tX}]$$

for all $t \in \mathbb{R}$ for which the expectation is well defined.

Theorem 3.4.3

Assume that M_X exists in a neighbourhood of 0, that is, there exists $\epsilon > 0$ such that for all $t \in (-\epsilon, \epsilon)$ we have $M_X(t) < \infty$. Then for all $k \in \mathbb{N}$ the kth moment of X exists, and

$$E[X^k] = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}$$

Proof. We have that $E[X^k] = \int_{-\infty}^{\infty} x^k f_X(x)\,dx$ for any continuous cummulative probability. On the other hand,

$$\begin{aligned}
\frac{d^k}{dt^k} M_X(t)\bigg|_{t=0} &= \frac{d^k}{dt^k} \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx\bigg|_{t=0} \\
&= \int_{-\infty}^{\infty} \frac{\partial^k}{\partial t^k} e^{tx} f_X(x)\,dx\bigg|_{t=0} \\
&= \int_{-\infty}^{\infty} x^k e^{tx} f_X(x)\,dx\bigg|_{t=0} \\
&= \int_{-\infty}^{\infty} x^k f_X(x)\,dx
\end{aligned}$$

□

Proposition 3.4.4

Assume that all expectations in the statement are well defined.

- For any $a, b \in \mathbb{R}$, $M_{aX+b}(t) = e^{tb} M_X(at)$
- If X, Y are independent, then $M_{X+Y}(t) = M_X(t) M_Y(t)$

Theorem 3.4.5

Let X, Y be two random variables. Assume that the moment generating functions of X, Y exists and are finite on an interval of the form $(-\epsilon, \epsilon)$. Assume further that $M_X(t) = M_Y(t)$ for all $t \in (-\epsilon, \epsilon)$. Then X, Y have the same distribution.

Theorem 3.4.6

Let X be a non-negative random variable whose expectation is well defined. We then have

$$P(X \geq x) \leq \frac{E(X)}{x}$$

Theorem 3.4.7

Let X be a random variable whose variance is well defined. Then

$$P(|X - E(X)| \geq x) \leq \frac{\operatorname{Var}(X)}{x^2}$$

for all $x > 0$

4 Convergence of Random Variables

4.1 Convergence

Definition 4.1.1: Convergence in Mean Square

We say that a sequence of random variables $X_1, X_2, \ldots$ converges in mean square to a random variable X if

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0$$

Definition 4.1.2: Convergence in Probability

We say that a sequence of random variables $X_1, X_2, \ldots$ converges in probability to a random variable X if for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

Theorem 4.1.3

Let $X_1, X_2, \ldots$ be a sequence of random variables, and X another random variable. If $X_n \to X$ in mean square as $n \to \infty$ then $X_n \to X$ in probability as $n \to \infty$.

Definition 4.1.4: Convergence in Distribution

We say that a sequence of random variables $X_1, X_2, \ldots$ converges in distribution to a random variable X if

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

for every x in the set $C = \{x \in \mathbb{R} : F_X \text{ is continuous at } x\}$.

Theorem 4.1.5

For any random variable X, the set of discontinuity points of F_X is countable.

Theorem 4.1.6

Let $X_1, X_2, \ldots$ be a sequence of random variables, and X another random variable. If $X_n \to X$ in probability, then $X_n \to X$ in distribution.

Theorem 4.1.7

Let $X_1, X_2, \ldots$ be a sequence of random variables such that $X_n \to c$ in distribution, where $c \in \mathbb{R}$, then X_n converges in probability to c.

Theorem 4.1.8: Law of large numbers in mean square

Let $X_1, X_2, \ldots$ be a sequence of independent random variable, each with mean μ and variance σ^2. Then

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \to \mu$$

in mean square.

Theorem 4.1.9: Weak law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of independent random variable, each with mean μ and variance $\sigma^2 \neq 0$. Then

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \to \mu$$

in probability.

4.2 Standardized Random Variables

Definition 4.2.1: Standardized Random Variables

Let X be a random variable with finite variance. We define the standardized version of X to be the random variable Z given by

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

Theorem 4.2.2: Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with mean μ and variance $\sigma^2 \neq 0$. Let $S_n = X_1 + \cdots + X_n$. Then the standardized version of S_n,

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution as $n \to \infty$ to a Gaussian random variable with mean 0 and variance 1. That is,

$$\lim_{n\to\infty} P(Z_n \leq x) = \lim_{n\to\infty} F_{Z_n}(x) = F_Y(y) = \int_{-\infty}^{x} -\frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

5 Stochastic Processes

5.1 Markov Chains

Definition 5.1.1: Stochastic Matrix

Let $P = (p_{i,j}) \in M_n(\mathbb{R})$ be a matrix. We say that P is a stochastic matrix if the following are true.

- $0 \leq p_{i,j} \leq 1$ for $1 \leq i, j \leq n$.
- For any fixed $1 \leq k \leq n$, we have

$$\sum_{j=1}^{n} p_{k,j} = 1$$

Definition 5.1.2: Markov Chain

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $I = \{1, \ldots, k\}$. Let $\{X_n : \Omega \to I \mid n \in \mathbb{N}\}$ be a sequence of random variables. Let $\lambda : I \to [0,1]$ be a probability distribution. Let $P = (p_{i,j})$ be a stochastic matrix. We say that $(X_n)_{n\geq 0}$ is a Markov chain with initial distribution λ and transition matrix P if the following are true.

- $\mathbb{P}(X_0 = i) = \lambda(i)$ for any $i \in I$.
- For any $i_0, \ldots, i_{n+1} \in I$, we have

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_k = i_k \text{ for } 0 \leq k \leq n) = \lambda(i_0) \cdot \prod_{j=0}^{n} p_{i_j, i_{j+1}}$$

In this case we say that $(X_n)_{n\geq 0}$ is Markov(λ, P).

In other words, Markov chains are a sequence of random variables where the next step does not depends on what states you visited previously, but only on the state now. There is an important property that makes studying Markov chains worth while.

Proposition 5.1.3

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $I = \{1, \ldots, k\}$. Let $\{X_n : \Omega \to I \mid n \in \mathbb{N}\}$ be a sequence of random variables. Let $\lambda : I \to [0,1]$ be a probability distribution. Then $(X_n)_{n\geq 0}$ is a Markov chain if and only if

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_k = i_k \text{ for } 0 \leq k \leq n) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n)$$

for any $i_0, \ldots, i_{n+1} \in I$. In this case, the transition matrix of the Markov chain is given by

$$P = \begin{pmatrix} \mathbb{P}(X_1 = 1 \mid X_0 = 1) & \cdots & \mathbb{P}(X_1 = k \mid X_0 = 1) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(X_1 = 1 \mid X_0 = k) & \cdots & \mathbb{P}(X_1 = k \mid X_0 = k) \end{pmatrix}$$

Theorem 5.1.4: Markov Property

Let $(X_n)_{n\geq 0}$ be a markov chain. Suppose that $X_m = E_m$ is given. Then $(X_{m+n})_{n\geq 0}$ is a Markov chain.

Notice that the transition probabbility $\mathbb{P}(X_{n+1} = j | X_n = i)$ still depedns on i, j, n. We further restrict out study of Markov chains to the following type.

Definition 5.1.5: Homogenous Markov Chains

We say that a Markov chain $(X_n)_{n\geq 0}$ is homogenous if

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i)$$

This means that homogenous Markov chains are time independent.

Lemma 5.1.6

The transition matrix P of a homogenous Markov chain is a Stochastic matrix.

Theorem 5.1.7

Let $(X_n)_{n\geq 0}$ be Markov. Then for all $m, n \geq 0$, we have

- $\mathbb{P}(X_n = j | X_0 = i) = \mathbb{P}(X_1 = j | X_0 = i)^n$
- $\mathbb{P}(X_n = j) = \sum_{i\in I} \mathbb{P}(X_0 = i)\mathbb{P}(X_n = j | X_0 = i)$

This theorem is saying that $\mathbb{P}(X_n = j | X_0 = i)$ is equal to the i, jth entry of the matrix P^n if P is the transition matrix. To find out the total probability of reaching the jth state at the $n$the step regardless of the starting point, you sum up the $\mathbb{P}(X_n = j | X_0 = i)$ multiplying the probability that you start at state i.

We often like to create new Markov chains from old (at least from exams). If we want to show that a squence $(X_n)_{n\geq 0}$ of random vriables is Markov, try and write $X_{n+1} = X_n +$ some term only related to $n+1$.

5.2 Communicating Classes

Definition 5.2.1: Talks and Communicates

Let i, j be states. We say that i talks to j which is $i \to j$ if there exists $n \in \mathbb{N}$ such that

$$\mathbb{P}(X_n = j | X_0 = i) > 0$$

We say that two states i and j communicate if $i \leftrightarrow j$.

Definition 5.2.2: Communicating Class

Let $C \subseteq I$. C is a communicating class if $\forall i, j \in C$, $i \leftrightarrow j$ and $\forall i \in C$ and $\forall k \in I \setminus C$, $i \not\to k$

Definition 5.2.3: Closed and Absorbing

We say that a class is closed if no states in the class talks to any states outside of that class. We say that a class is absorbing if it forms a closed class by itself.

Definition 5.2.4: Irreducible Markov Chains

A Markov chain is said to be irreducible if for all $i, j \in I$, $i \leftrightarrow j$.

5.3 Hitting Times

Definition 5.3.1: Stopping Times

Let $(X_n)_{n\geq 0}$ be Markov. Then a random variable

$$T : \Omega \to \mathbb{N} \cup \{\infty\}$$

is a stopping time if for all $n \geq 0$, the event $\{T = n\}$ depends only on $X_0, \ldots, X_n$.

Definition 5.3.2: Hitting Times

Let $(X_n)_{n\geq 0}$ be Markov. The hitting time of a subset $A \subseteq I$ is a random variable

$$H_A : \Omega \to \mathbb{N} \cup \{\infty\}$$

defined by

$$H_A(\omega) = \inf\{n \geq 0 | X_n(\omega) \in A\}$$

We use

$$h_i^A = \mathbb{P}(H^A < \infty | X_0 = i)$$

to denote the corresponding probability. We use

$$k_i^A = E_i[H_A] = \sum_{k=0}^{\infty} k\mathbb{P}(H_A = k | X_0 = i) + \infty \mathbb{P}(H_A = \infty | X_0 = i)$$

to denote the expectation.

Proposition 5.3.3

We can find the probabilities of all hitting times by solving the system of linear equations

$$\begin{cases} \mathbb{P}(H_A < \infty | X_0 = i) = 1 & i \in A \\ \mathbb{P}(H_A < \infty | X_0 = i) = \sum_{j \in I} \mathbb{P}(X_1 = j | X_0 = i)\mathbb{P}(H_A < \infty | X_0 = j) & x \notin A \end{cases}$$

Proposition 5.3.4

We can find the expected number of hitting times by solving the system of linear equations

$$\begin{cases} E_i[H_A] = 0 & i \in A \\ E_i[H_A] = 1 + \sum_{j \notin A} \mathbb{P}(X_1 = j | X_0 = i) E_j[H_A] & i \notin A \end{cases}$$

5.4 Strong Markov Property

Theorem 5.4.1: Strong Markov Property

Let $(X_n)_{n\geq 0}$ be Markov(λ, P) and T a stopping time of $(X_n)_{n\geq 0}$. Then conditional on both $\{X_T = i\}$ and $\{T < \infty\}$, we have $(X_{T+n})_{n\geq 0}$ is Markov(δ_i, P) and independent of $X_0, \ldots, X_T$.

5.5 Recurrence and Transcience

Definition 5.5.1: Recurrence and Transcience

A state $i \in I$ is recurrent if $\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 1$. A state $i \in I$ is transient if $\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 0$.

Definition 5.5.2: k-th Passage Time

The first passage time of state i is a stopping time such that

$$T_i(\omega) = \inf\{n \geq 1 | X_n(\omega) = i\}$$

The k-th passage time is a stopping time such that

$$T_i^{(k)}(\omega) = \inf\{n \geq T_i^{(k-1)} | X_n(\omega) = i\}$$

The k-th excursion time is defined by

$$S_i^{(k)} = \begin{cases} T_i^{(k)} - T_i^{(k-1)} & \text{if } T_i^{(k-1)} < \infty \\ \infty & \text{otherwise} \end{cases}$$

Intuitively, $T_i^{(k)}$ outputs the time it takes for the Markov chain to reach state i for the kth time and S_i^k outputs the number of steps taken between consecutive visits.

Lemma 5.5.3

For $k = 2, 3, \ldots$ and if $T_i^{(k-1)} < \infty$, then $S_i^{(k)}$ is independent of $\{X_m : m \leq T_i^{(k-1)}\}$ and

$$\mathbb{P}(S_i^{(k)} = n | T_i^{(k-1)} < \infty) = \mathbb{P}_i(T_i = n)$$

Definition 5.5.4: Visit Counting Function

Let $i \in I$ be a state. Define

$$V_i = \sum_{k=0}^{\infty} 1_{\{X_n = i\}}$$

the number of visits ever to state i.

Lemma 5.5.5

If V_i counts the number of vists to state i, then

$$E_i(V_i) = \sum_{k=0}^{\infty} P_{ii}^{(k)}$$

Lemma 5.5.6

For $k = 0, 1, 2, \ldots$, we have

$$\mathbb{P}_i(V_i > k) = (\mathbb{P}_i(T_i < \infty))^k$$

This lemma means that V_i asserts a geometric distribution.

Theorem 5.5.7: Characteristic of Recurrent and Transient States

Let $(X_n)_{n \geq 0}$ be a Markov chain. Let T_i be the first passage time of state i. If $\mathbb{P}_i(T_i < \infty) = 1$, then i is recurrent and $\sum_{k=1}^{\infty} \mathbb{P}(X_1 = k | X_0 = k)^n = \infty$. If $\mathbb{P}_i(T_i < \infty) < 1$, then i is transient

and $\sum_{k=1}^{\infty} \mathbb{P}(X_1 = k | X_0 = k)^n < \infty$.

Theorem 5.5.8

Let C be a communicating class. Then either all states in C are transient or all are recurrent.

Theorem 5.5.9

Let $C \subset I$ be a class of a Markov chain. Then
- Every recurrent class is closed
- Every finite closed class is recurrent

Theorem 5.5.10

If P is irreducible and recurrent then for all $j \in I$,

$$\mathbb{P}(T_j < \infty) = 1$$

Definition 5.5.11: Component Independent Simple Random Walk on $\mathbb{Z}^d$

A component independent simple random walk on $\mathbb{Z}^d$ for $d \in \mathbb{N}$ is defined as

$$X_{n+1} = X_n + Z_{n+1}$$

where $Z_{n+1} = (Z_{n+1}^1, \ldots, Z_{n+1}^d)$ with

$$Z_m^j = \begin{cases} 1 & \text{with probability } p_j \\ -1 & \text{with probabilty } q_j \end{cases}$$

where Z_m^j are independent for all j, m.

Proposition 5.5.12

Let $(X_n)_{n \geq 0}$ be a CISRW on $\mathbb{Z}^d$. If there exists $j \in \{1, \ldots, d\}$ such that $p_j \neq \frac{1}{2}$ then (X_n) is transient.

Proposition 5.5.13

Let $(X_n)_{n \geq 0}$ be a CISRW on $\mathbb{Z}^d$ and $p_j = q_j = \frac{1}{2}$ for all j. Then
- If $d \leq 2$ then all states are recurrent
- If $d \geq 3$ then all states are transient

5.6 Branching Process

Definition 5.6.1: Branching Process

Let $(X_n)_{n\geq 0}$ be the number of individuals in a population at time n where $X_n \in \{0\} \cup \mathbb{N}$. At every time step each individual in the population gives birth to a random number of off-spring. Thus X_{n+1} is defined to be

$$X_{n+1} = \sum_{k=1}^{X_n} Z_k^n$$

where $Z_k^n \sim Z$ and $\mathbb{P}(Z \geq 0) = 1$. Then $(X_n)_{n\geq 0}$ is said to be a branching process.

Proposition 5.6.2

Define $G(s) = E[s^Z]$ and $F_{n+1}(s) = E[S^{X_{n+1}} | X_0 = 1]$ for $s \in (0, 1)$. Then

$$F_n(s) = G(F_{n-1}(s))$$

Proposition 5.6.3

For a braching process $(X_n)_{n\geq 0}$ with off-spring distribution Z where $E[Z] = \mu$, we have

$$E[X_n] = \mu^n$$

Theorem 5.6.4: Extinction Probability

The extinction probability is the smallest non-negative root of $G(\alpha) = \alpha$.

Theorem 5.6.5

Suppose that $G(0) > 0$, $\mu = E[Z] = G'(1)$. Then $\mu \leq 1$ implies certain extiction. $\mu > 1$ implies uncertain extinction.

5.7 Invariant Distributions on a Markov Chain

Definition 5.7.1: Invariant Distribution

A measure $\lambda = (\lambda_i : i \in I)$ with non-negative entries is said to be an invariant distribution if

- $\lambda P = P$
- $\pi_i \in [0, 1]$ for all i
- $\sum_{i\in I} \pi_i = 1$

The following theorem shows that after applying a Markov process to an invariant distribution, the new distribution will be the same as the old one. Notice here that right multiplication of a distribution gives the next step on the Markov chain instead of the usual left multiplication.

Theorem 5.7.2

Let $(X_n)_{n\geq 0}$ be Markov(π, P), where π is an invariant distribution for P. Then $\forall m \geq 0$, $(X_{n+m})_{n\geq 0}$ is Markov(π, P).

Proof. By the Markov property, the new sequence will be Markov. But we do not yet know

its distribution. We have that

$$\begin{aligned}\mathbb{P}(X_m = i) &= (\pi P^m)_i && \text{(this denotes the } i\text{th entry of } P)\\ &= (\pi P P^{m-1})_i \\ &= (\pi P^{m-1})_i\end{aligned}$$

Thus by induction, we see that $\mathbb{P}(X_m = i) = \pi_i$ thus we are done. □

Theorem 5.7.3

Let I be finite. Suppose that for some $i \in I$,

$$P_{ij}^n \to \pi_j$$

for all $j \in I$. Then $\pi = (\pi_j : j \in I)$ is an invariant distribution.

Let us look at an example.

Example 5.7.4: Gene Mutation

Recall the gene mutation example that has transition matrix $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$. Recall that

$$P^n = \begin{pmatrix} \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n & \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n \\ ? & ? \end{pmatrix}$$

As $n \to \infty$, observe that the first entry tends to $\frac{\beta}{\alpha+\beta}$ the second entry tends to $\frac{\alpha}{\alpha+\beta}$. This means that $\pi = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$ is an invariant distribution for the gene mutation model.

We can also find the invariant distribution directly: Let $\pi = (\pi_1, \pi_2)$ be an invariant distribution for P. We have that

$$\begin{aligned} \pi P &= \pi \\ \begin{pmatrix} \pi_1(1-\alpha) + \pi_2\beta & \pi_1\alpha + \pi_2(1-\beta) \end{pmatrix} &= \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \\ \begin{pmatrix} \pi_2\beta - \pi_1\alpha & \pi_1\alpha - \pi_2\beta \end{pmatrix} &= 0 \end{aligned}$$

Notice that these two equations mean the same thing. (In general if you have an $n \times n$ transition matrix, only the first $n-1$ equations will be useful and the last one will be redundant.) This is why we need to use the fact that $\pi_1 + \pi_2 = 1$ as our final equation.

Now solving it, we have that $\pi = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$ which is the exact same answer as the one given in the first method.

A third way to think about it is that notice that $P^T\pi^T = \pi^T$ is the equation that we are attempting to solve. Recall that 1 is necessarily an eigenvector of a transtition matrix which means that this equation really is just a question of finding eigenvectors from the eigenvalue 1.

The above examples gives a lot of practical information when calculating the invariant distribution of a transition matrix.

Definition 5.7.5: Expected Time Spent

Define the expected time spent in state i in between visits to state k by

$$\gamma_i^k = E_k\left(\sum_{n=0}^{T_k-1} \mathcal{X}_{X_n=i}\right)$$

This uses k as a reference, and as we keep going back to k, this records how much time we have spent in i in the process of leaving and returning to k.

Theorem 5.7.6

Let P be irreducible and recurrent. Then

- $\gamma_k^k = 1$
- $\gamma^k = (\gamma_i^k : i \in I)$ is an invariant measure satisfying $\gamma^k P = \gamma^k$
- $0 < \gamma_i^k < \infty \ \forall i \in I$

Theorem 5.7.7

Let P be irreducible and λ and invariant measure and $\lambda_k = 1$ for some k. Then $\lambda \geq \gamma^k$. If we also have that P is recurrent, then $\lambda = \gamma^k$

This theorem assesrts that as long as P is irreducible and recurrent, then any invariant measure is exactly equal to the γ^k we defined, which means that there is really only one invariant measure up to rescaling.

Definition 5.7.8: Positive Recurrent

A state $i \in I$ is positive recurrent if it is recurrent and

$$E_i[T_i] < \infty$$

A state which is not positive recurrent but is recurrent is called null recurrent.

Theorem 5.7.9

Let P be irreducible. Then the following are equivalent.

- Every state is positive recurrent
- Some state i is positive recurrent
- P has an invariant distribution π. More over, $\pi_i = \frac{1}{E_i[T_i]}$