A BRIEF INTRODUCTION TO STOCHASTIC PROCESSES

Introduction and Review of Probability Theory

WANG SHAOCHEN

Last update: June 11, 2023, draft version

about 6∼8 class hours

School of Mathematics

South China University of Technology

✉mascwang@scut.edu.cn

References

- **Some (elementary) reference books are**
  - *Essentials of stochastic processes*, **R. Durrett, 3rd edition, Springer, 2016.** ✓✓✓ 主 参 书， 中
  - *An introduction to stochastic modeling*, **M.A. Pinsky and S. Karlin, 4th edition, Academic Press, 2011.** ✓✓✓
  - *Introduction to probability models*, **S.M. Ross, 12th edition, Academic Press, 2021.** ✓✓✓ 中
  - *Introduction to stochastic processes with R*, **R.P. Dobrow, Wiley, 2016.** ✓✓
  - ，何书元，北京大学出 ，**2008.** ✓✓
  - ，劳 **(S.M. Ross)**， 工业出 (中 ， 二)，**1995.** ✓✓

Contents of This Chapter

1. Stochastic Processes: Definitions and Examples
2. Review of Probability Theory
   - Probability Space
   - Conditional Probability
   - Random Variables and Random Vectors
   - (Conditional) Expectations
   - Characteristic Functions and Limits Theorems

We demand rigidly defined areas of doubt and uncertainty!

-Douglas Adams

Deterministic vs. Stochastic Models: Motivations

- **Why we use stochastic or random models instead of deterministic models? We begin from an concrete example.**

Example 1 (Bacteria Population Growth Model). **Let** yK B i2 2i g °

Definitions of Stochastic Processes

- **Roughly speaking, stochastic process is a collection of RVs.**

Definition 1. **Let $X = \{X_\alpha, \alpha \in \mathbb{T}\}$ be a collection of random variables which defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and each takes values in measurable space $(S, \mathfrak{S})$, where $\mathbb{T}$ is some (possible abstract) index set.**

(1) **If $\mathbb{T}$ only takes some discrete values, say $\{0, 1, 2, \cdots\}$, then we call X is a discrete time stochastic process.**

(2) **If $\mathbb{T} = [T_1, T_2]$, then we call X is a continuous time stochastic process, where $0 \le T_1 < T_2 \le +\infty$.**

(3) **Some other cases, for example, multivariate parameters stochastic process, stochastic fields and so on.**

Several Remarks

- **In this course, we only consider the cases in (1) and (2) and we may also use the parentheses notation $\{X(t), t \geq 0\}$ in the continuous time stochastic processes case, for example, Poisson Process $\{N(t), t \geq 0\}$ and Brownian Motion $\{B(t), t \geq 0\}$.**
- **Usually, the index set $\mathbb{T}$ is interpreted as the evolution of time. In the following, S will be known as the state space of the stochastic process, $\mathfrak{S}$ is the σ-algebra generated by some subset of S.**
- **Most of cases, $S = \mathbf{R}^d, d \geq 1, \mathfrak{S} = \mathscr{B}(S)$ or its subset. For example, for Markov Chain and Poisson Process $S = \{0, 1, 2, \cdots\}$ or its subset, and $S = \mathbf{R}$ for 1-dimensional Brownian Motion.**
- **You can ignore the notation $\mathfrak{S}$ without any risk!**

Sample Path

- **The observed samples of $X = \{X_\alpha, \alpha \in \mathbb{T}\}$, as a function of α, is called a sample path (or realization) of X. For example in the continuous time case,**

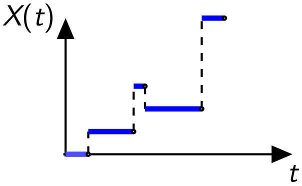


Figure: A sample path of $X(t)$

- **Sample paths can be rather complicated and disordered. The (invariant) law behind these superficiality is our main focus.**

Two Important Concepts

- **We always assume the subscript of a stochastic process lies in the domain of $\mathbb{T}$ following.**

Definition 2. Let X be a discrete or continuous time real-valued, (i.e., $S = \mathbf{R}^d, d \geq 1$ or its subset) stochastic process, we call it has

(1) **independent increments, if for any $s_1 \leq s_2 \leq \cdots \leq s_n$,**

$$X_{s_2} - X_{s_1}, X_{s_3} - X_{s_2}, \cdots, X_{s_n} - X_{s_{n-1}}$$

are independent random variables;

(2) **stationary increments, if for any $s < t$ and $h > 0$, $X_t - X_s$ has the same distribution with $X_{t+h} - X_{s+h}$.**

- **One simplest example is**

Example 3 (Simple (Symmetric) Random Walk). **Assume $X_0 = x \in \mathbb{N}$ and $X_i, i \geq 1$ are iid. (here and the following, "iid." is abbreviation of independent and identically distributed) random variables, with common distribution**

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2.$$

- **Define**

$$S_n = X_0 + X_1 + \cdots + X_n,$$

then we call $\{S_n, n \geq 0\}$ is a simple (symmetric) random walk (starts from x) on the real line.

- S_n **can be interpreted as the position of the random walk (starts form x) at the n-th step.**

- $\{S_n, n \geq 0\}$ **can be used to model many interesting phenomenon, for example the gambling game.**

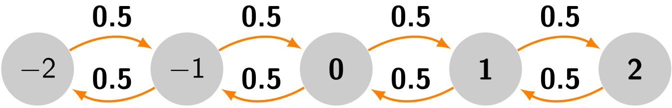


Figure: Simple symmetric random walk

- **Usually, we want to know:**
  - **What is the limiting behaviors (macro property) of S_n? (Law of large numbers, central limit theorem and ...)**
  - **Whether S_n hits $b \in \mathbb{Z}$ in some finite time with probability one or less than one?**
  - **What is expected time when we first observe $S_n = b$?**

Example 4 (Simple Random Walk with Absorbing Walls). **Under the setting of above example with $x=0$ and let a,b be two positive integers, we denote**

$$\tau_d = \inf\left\{n \geq 1 : S_n = d\right\}, \quad \tau := \tau_{-a} \wedge \tau_b := \min\{\tau_{-a}, \tau_b\},$$

where $d=-a$ or b. Define $Y_n = S_{n\wedge\tau}$, then $\{Y_n, n \geq 0\}$ is called the simple random walk with absorbing walls $\{-a,b\}$.

- **The integers a,b can be interpreted as the capital of gambler and gambling house respectively. Generally, $b \gg a$ (means b is much larger than a), the interesting thing is we can prove**

$$\mathbb{P}(\tau < +\infty) = 1 \text{ and } \mathbb{P}(\tau_{-a} < \tau_b) = \frac{b}{a+b}.$$

- **More advanced topics will be further studied later.**

Branching Process (分支过程)

- In the mid 19th century several aristocratic families in Victorian England realized that their family names could become extinct.
- Francis Galton posed the following question (1873, *Educational Times*):

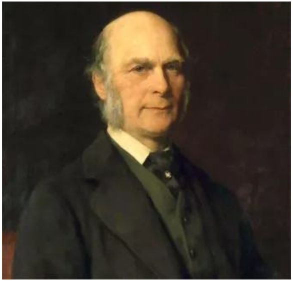


Figure: Francis Galton[a]

[a] 没错！就是你想到的那个高尔顿

How many male children (on average) must each generation of a family have in order for the family name to continue in perpetuity?

Example 5 (Inventory Model, $(s, \mathscr{S})$ Control Policy). **Let X_n be the amount of products at the end of n-th day and D_{n+1} be the demand on $(n+1)$-th day. Assume $D_n, n \geq 1$ are iid. random variables and are independent with $\{X_n, n \geq 1\}$.**

Suppose $0 < s < X_0 \leq \mathscr{S}$, an $(s, \mathscr{S})$ control policy means that

$$X_{n+1} = \begin{cases} (X_n - D_{n+1})_+, & \textbf{if } X_n > s; \\ (\mathscr{S} - D_{n+1})_+, & \textbf{if } X_n \leq s, \end{cases}$$

for $n \geq 0$, where x_+ means $\max\{x, 0\}$.

Example 6 (Continued). **For example, an electronics store sells a game video and uses above inventory control policy with $s = 1, \mathscr{S} = 5$. Assume D_n has following distribution**

$$\mathbb{P}(D_n = 0) = 0.3, \quad \mathbb{P}(D_n = 1) = 0.4$$
$$\mathbb{P}(D_n = 2) = 0.2, \quad \mathbb{P}(D_n = 3) = 0.1.$$

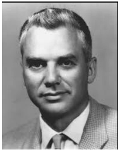
The interesting question is: suppose we make \$12 profit on each unit sold but it costs \$2 per day to store one.

- **What is the long-run profit per day of this inventory policy?**
- **Further, can we find the optimal choice of $(s, \mathscr{S})$ so that makes the largest profit per day in long run?**

- **Above queries will be answered using Markov Chain theory.**

Martingale Process

- **The word martingale ( ) comes from the French city of Martigues, whose inhabitants were said to know about a betting strategy that guaranteed a profit.**
- **The martingale theory establishes that it is not quite possible... (under reasonable assumptions). Much of the original development of the martingale theory was done by Joe Doob. Part of his initial motivation was to show the impossibility of successful betting strategies.**
- **Joseph Leo Doob (1910-2004) was an American mathematician, regarded as the father of the theory of martingales.**



The "Martingale" strategy (or James Bond's strategy)

- **Suppose the game is fair, that is, $p=1/2$, our strategy is:**
  - $\rightsquigarrow$ **bet 1 on the first game; stop if you win**
  - $\rightsquigarrow$ **if not, bet 2 on the second game; stop if you win**
  - $\rightsquigarrow$ **...**
  - $\rightsquigarrow$ **stop betting if you win, if not double your bet**
  - $\rightsquigarrow$ **...**
  - $\rightsquigarrow$ **keep doubling your bet until you eventually win.**
- **The duration time of game follows a Geometric distribution, so this game will be ended in finite time with probability one and more surprisingly, you are sure to end up richer than when you started! So where is the trick? The martingale theory will tell you that you can not beat an unfavorable game!!!**

- **Martingale is one of most important concept in financial mathematics, a continuous time martingale example is**

Example 7 (Brownian Motion, BM). **A one dimensional real-valued stochastic process $B = \{B(t), t \geq 0\}$ is a Brownian motion if it satisfies the following properties**

(1) $B(0) = 0$ **almost surely;**

(2) B **has independent increments, and for** $s < t$**, the increment** $B(t) - B(s)$ **has a normal distribution with mean** 0 **and variance** $t - s$**;**

(3) **almost all the sample paths of** B **are continuous.**

Brief History about BM

- **1827, Brown; 1905, Einstein; 1921, Wiener.**
- **BM is now widely used in financial mathematics.**

Review of Probability Theory

"*The probable is what usually happens*"

Aristotle (384-322 BCE)

Review of Probability Theory

- **A random experiment is an experiment which outcomes can not be predicted. Let Ω be the collection of all possible outcomes, which is called as sample space.**
- **When $\#\Omega$ is finite, all subsets of Ω are events.**
- **The operations between events are same as sets, which we list them in following exercise and one can check it easily.**

Exercise 1. ☕ **(1) Recall the definitions and operation rules between sets, such as A^c (or $\overline{A}$), $A \cup B$, $A \cap B$, $A - B$.**
(2) Prove the De Morgan's laws

$$\left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c, \qquad \left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c.$$

- **When $\#\Omega$ is finite, we can define the classical probability on Ω which making all subsets of Ω are events. For each event A, we can assign a "probability" $\mathbb{P}(A)$, it should satisfy**

(1) $\mathbb{P}(\Omega) = 1, \ \mathbb{P}(A) \geq 0$ **for all** A**;**

(2) **when** $A \cap B = \emptyset$**, then**

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

-  **But this will be problematic when $\#\Omega$ is infinite. For example, if $\Omega = [0,1]$, we can not assign a probability to each $\omega \in \Omega$ (even we cannot assign prob. to all subsets of $[0,1]$).**
- **We need to introduce the axiomatization construction of probability (due to the great contribution of Kolmogorov).**

Algebra and σ-algebra

Probability and Probability Space

Definition 4 ((Probability) Measure). Let $(\Omega, \mathscr{F})$ be a measurable space. A **measure** on $\mathscr{F}$ is some set function ν which has following properties:

(1) Nonnegativity, for any $A \in \mathscr{F}$, $\nu(A) \in [0, +\infty]$.

(2) $\nu(\emptyset) = 0$.

(3) If $A_i \in \mathscr{F}, i = 1, 2, \cdots$ and are disjoint, then

$$\nu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \nu(A_j).$$

- Usually ν is a probability measure, i.e., $\boxed{\nu(\Omega) = 1}$. We use $\mathbb{P}$ instead of ν. The triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a **probability space.**

Proposition 1. Assume $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space.

(1) **(Monotonicity) If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.**

(2) **(Subadditivity or Boole's inequality)**

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) \leq \{\searrow\sim\searrow\cdot\searrow$$

"

D=

$(\Omega[[(;)]$

Td[(()]TJ8m0 Do /F610 10.:=Tf 10.9 0 Td[(P)]TJ /3 g 000.909 Tf 7.88 18.66 Td[( )]TJ /F4 7.9

D=

One Application: Extinction Probability

Example 8 (Extinction Probability). **Let $X_n, n \geq 0$ be the number of population in the n-th generation of some animal, then $\{X_n = 0\}$ implies $\{X_{n+1} = 0\}$, thus by the continuity of probability, we have**

$$\mathbb{P}(\textbf{extinction of animal}) = \mathbb{P}\left(\bigcup_{n=0}^{\infty}\{X_n = 0\}\right)$$

$$= \mathbb{P}\left(\lim_{n\to\infty}\{X_n = 0\}\right) = \lim_{n\to\infty}\mathbb{P}(X_n = 0).$$

- **The first equality holds since the animal extinction if and only if there exists some n so that the event $\{X_n = 0\}$ happened.**
- **The term $\mathbb{P}(X_n = 0)$ will be considered in Branching Process.**

Conditional Probability

- **For any event A such that $\mathbb{P}(A) > 0$, the conditional probability of B given A is defined as**

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}}.$$

- **This formula is known as conditional probability formula. A direct corollary of conditional probability formula is the multiplication rule, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A)$.**

Exercise 2. ☕ **(1) Prove the generalized multiplication rule,**

$$\mathbb{P}(A_1A_2\cdots A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)\cdots\mathbb{P}(A_n|A_1\cdots A_{n-1}).$$

(2) For set A with $\mathbb{P}(A) > 0$, prove $\mathbb{P}_A(\cdot) := \mathbb{P}(\cdot|A)$ is a probability.

Definition 6 (Independence of Events). **Let $A_i, i = 1, 2, \cdots, n$ be a sequence of events. We say they are**

(1) **pairwise independent, if for any $i, j \in \{1, 2, \cdots, n\}$, A_i and A_j are independent;**

(2) **mutual independent, if for any $1 \le i_1 < i_2 < \cdots < i_k \le n$,**

$$\mathbb{P}(A_{i_1} A_{i_2} \cdots A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

(3) **A sequence of events $A_n, n \ge 1$ are called pairwise (mutual) independent, if for any $i_1, i_2, \cdots, i_n$, the events $A_{i_1}, A_{i_2}, \cdots, A_{i_n}$ are pairwise (mutual) independent.**

- **We remark that the mutual independence implies pairwise independence. But the converse may not hold generally.**

Random Variables and Random Vectors

- **A random variable X is a Borel measurable function from $(\Omega, \mathscr{F})$ to $(\mathbf{R}, \mathscr{B}(\mathbf{R}))$, i.e., for any $A \in \mathscr{B}(\mathbf{R})$,**

$$X^{-1}(A) := \{\omega | X(\omega) \in A\} \in \mathscr{F}.$$

- **Roughly speaking, a vector is called random vector if its components are random variables. (A d-dimensional random vector is a Borel measurable function from Ω to $\mathbf{R}^d$.)**

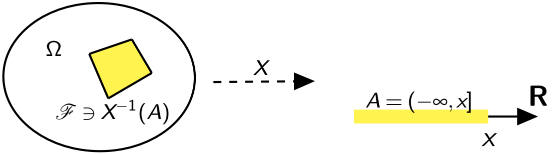


Figure: Borel measurable mapping from $(\Omega, \mathscr{F})$ to $(\mathbf{R}, \mathscr{B}(\mathbf{R}))$

- **As a simplest example, suppose we flip a coin with sample space $\Omega = \{H, T\}$ and $\mathscr{F} = \{\{H\}, \{T\}, \emptyset, \Omega\}$. We define**

$$X(\omega) = \begin{cases} 1, & \textbf{if } \omega = H\textbf{;} \\ 0, & \textbf{if } \omega = T\textbf{.} \end{cases}$$

Then clearly X is a random variable (Exercise!!!).

- **Now if we flip the coin n times and define $X_i, i \geq 1$ similarly as above, i.e., $X_i = 1$ if the i-th flip result is H and 0 otherwise. Then $X_1 + X_2 + \cdots + X_n$ means the total times of H in total n flips. The operations between random variables will be much easier than events.**

- **It can be proved that the usual arithmetic operations between random variables or compositions with Borel measurable functions are still random variables, for example, $X + Y$, XY, X^2, $\sin(X)$ and so on.**

Distribution Function

- **Usually, random variable (or vector) X can be described by its distribution, which is defined by**

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(\{\omega : X(\omega) \in A\}) \\ &= \mathbb{P}(X^{-1}(A)), \quad A \in \mathscr{B}(\mathbf{R}).\end{aligned}$$

- **For example, in above coin toss example,** $\mathbb{P}(X=1)=\mathbb{P}(\{H\})=1/2$ **if the coin is fair. Particularly, for** $A=(-\infty,x]$, $x \in \mathbf{R}$**,**

$$\mathbb{P}(X \in A) = \mathbb{P}(X \le x) =: F_X(x),$$

which is known as the cumulative distribution function (abbr. cdf.) or simply as distribution function. We have $F(-\infty)=0$, $F(+\infty)=1$ **and** $F(x)$ **is non-decreasing and right continuous.**

- **There are two special and important classes of RVs.**
  - **<u>Discrete case</u>, X only takes finite or countably many values, we can describe it by probability mass function, i.e.,**

$$p_i := \mathbb{P}(X = x_i) > 0, \quad i = 1, 2, \cdots.$$

  - **<u>Continuous case</u>, which there exists some real-valued measurable function $f \geq 0$ defined on R, called probability density function so that for all $a \leq b$,**

$$\mathbb{P}(a < X \leq b) = \int_a^b f(x)\mathrm{d}x.$$

- **We remark that if X has probability density function f, we can prove for any $A \in \mathscr{B}(\mathbf{R})$,**

$$\boxed{\mathbb{P}(X \in A) = \int_A f(x)\mathrm{d}x}.$$

- **Similarly, for random vector $\mathbf{X} := (X_1, X_2, \cdots, X_n)$, we also have two special classes.**
  - **Discrete case, its joint probability mass function**

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

  - **Continuous case, for any Borel set $A \subset \mathbf{R}^n$,**

$$\mathbb{P}\Big((X_1, X_2, \cdots, X_n) \in A\Big) = \iint \cdots \int_A f(x_1, \cdots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

    and $f(x_1, x_2, \cdots, x_n)$ is called the joint probability density function.
- **Usually, we often deal with the case $n = 2$. The probability mass (density) function of each random component can be recovered from the joint probability mass (density) function.**

Example 10. **Marginal probability mass (density) functions.**

(1) **Discrete case**

$$\mathbb{P}(X=x)=\sum_y \mathbb{P}(X=x,Y=y),\quad \mathbb{P}(Y=y)=\sum_x \mathbb{P}(X=x,Y=y).$$

(2) **Continuous case**

$$f_X(x)=\int f(x,y)\mathrm{d}y,\quad f_Y(y)=\int f(x,y)\mathrm{d}x.$$

Definition 7 (Independence of Random Variables). For two random variables X, Y, we call they are independent, if for any two Borel sets A, B, we have

$$\mathbb{P}(X\in A, Y\in B)=\mathbb{P}(X\in A)\mathbb{P}(Y\in B).$$

- X, Y **are independent if and only if** $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.

Example 11. X, Y **are independent if and only if for all possible values of** x, y **so that**

(1) **Discrete case**

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y);$$

(2) **Continuous case**

$$f(x,y) = f_X(x)f_Y(y);$$

(3) **Mixture case, i.e.,** X **is discrete and** Y **is continuous,**

$$\mathbb{P}(X = x_i, Y \leq y) = \mathbb{P}(X = x_i)\mathbb{P}(Y \leq y).$$

(Conditional) Expectations

- **Expectation describes the average value of random variable.**

Definition 8. (1) If X is a discrete random variable so that $\sum_k |x_k| \mathbb{P}(X = x_k) < +\infty$, then its expectation is defined as

$$\mathbb{E}[X] = \sum_k x_k \, \mathbb{P}(X = x_k) \, .$$

(2) Similarly, if X is a continuous random variable so that $\int |x| f_X(x) \mathrm{d}x < \infty$, then its expectation is defined as

$$\mathbb{E}[X] = \int x \, f_X(x) \, \mathrm{d}x.$$

- **Why we use the word "similarly" in above definition (2)?**

- **For general random variable X, its expectation is defined as**

$$\mathbb{E}[X] = \int x \mathrm{d} F_X(x)$$

provided the integral exists. The integral on the right hand side is interpreted as Stieltjes integral. Here are two special cases

(1) **If $F_X(x)$ is absolutely continuous with respect to the Lebesgue measure, then**

$$\int x \mathrm{d} F_X(x) = \int x \frac{\mathrm{d} F_X(x)}{\mathrm{d} x} \mathrm{d} x,$$

in this case X is a continuous r.v. with pdf. $\frac{\mathrm{d} F_X(x)}{\mathrm{d} x}$.

(2) **If $F_X(x)$ is a step function with jumps at $x_j, j \geq 1$, then**

$$\int x \mathrm{d} F_X(x) = \sum_j x_j [F_X(x_j) - F_X(x_j-)],$$

in this case X is a discrete r.v. with pmf. $\mathbb{P}(X = x_j) = F_X(x_j) - F_X(x_j-)$, $j \geq 1$.

- **Next conclusion is vital important, it gives simple computation formula for functions of random variables (vectors).**

om

Example 12. **Assume X, Y are independent with common distribution $N(0,1)$, please find $\mathbb{E}[|X|]$ and $\mathbb{E}[\max\{X, Y\}]$.**

- **The variance of X is defined as**

$$\mathrm{Var}[X] = \int (x - \mathbb{E}[X])^2 \mathrm{d}F_X(x) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

provided the integral exists. Particularly, we have

$$\mathrm{Var}[X] = \int (x-\mu)^2 f_X(x)\mathrm{d}x, \quad \mathrm{Var}[X] = \sum_k (x_k - \mu)^2 \mathbb{P}(X = x_k)$$

in continuous and discrete cases for X, where $\mu = \mathbb{E}[X]$.

- **For two random variables X, Y, its covariance is defined as**

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

If $\mathrm{Cov}(X, Y) = 0$, we call X, Y are uncorrelated.

Fubini's Theorem

- **In probability theory, we often meet the case of whether we can interchange the orders of integrations. Next theorem gives some sufficient conditions.**

Theorem 5 (Fubini's Theorem). **Let f be a Borel function on $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \times \mathscr{F}_2)$ and ν_1, ν_2 be two σ-finite measure on each of it.**[a] **Suppose either that $f \geq 0$ or f is integrable w.r.t. $\nu_1 \times \nu_2$, then**

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) \mathrm{d}(\nu_1 \times \nu_2) = \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2) \mathrm{d}\nu_1 \right] \mathrm{d}\nu_2.$$

[a] Ignore the concepts here, we only need to know how to use this theorem.

- **Two special cases are**

Example 13. **Let $\Omega_1 = \Omega_2 = \{0, 1, 2, \cdots\}$ and ν_1, ν_2 be the counting measures. A function f on $\Omega_1 \times \Omega_2$ defines a double sequence. If $f \geq 0$ or $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |f(i,j)| < \infty$, then**

$$\int f \,\mathrm{d}\nu_1 \times \nu_2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(i,j) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} f(i,j).$$

Example 14. **Suppose (X, Y) is a random vector with density function $h(x,y)$, then for any $f \geq 0$ or $\mathbb{E}[|f(X,Y)|] < +\infty$, then we have**

$$\begin{aligned}\mathbb{E}[f(X,Y)] = \iint f(x,y)h(x,y)\mathrm{d}x\mathrm{d}y &= \int \left(\int f(x,y)h(x,y)\mathrm{d}x \right) \mathrm{d}y \\ &= \int \left(\int f(x,y)h(x,y)\mathrm{d}y \right) \mathrm{d}x.\end{aligned}$$

- **As applications of Fubini's theorem, next two results give simple but important computational formulas for expectations.**

Theorem 6 (Tail Type Moments Formulas). **(1) If $X \geq 0$ is integer-valued, then**

$$\boxed{\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k)} = \mathbb{P}(X \geq 1) + \mathbb{P}(X \geq 2) + \cdots$$

$$= \sum_{k=0}^{\infty} \mathbb{P}(X > k) = \mathbb{P}(X > 0) + \mathbb{P}(X > 1) + \cdots.$$

(2) Assume $X \geq 0$, for any $p > 0$, we have

$$\mathbb{E}[X^p] = \int_0^{\infty} p x^{p-1} \mathbb{P}(X > x) \mathrm{d}x.$$

- **Particularly, we have $\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x) \mathrm{d}x$.**

Theorem 7 (Serval Useful Inequalities). **(1) Markov inequality. For any $\delta > 0$,**

$$\mathbb{P}(|X| \geq \delta) \leq \frac{1}{\delta^\alpha}\mathbb{E}[|X|^\alpha], \quad \alpha > 0.$$

- **Particularly, we have the Chebyshev inequality,**

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \delta) \leq \frac{1}{\delta^2}\mathrm{Var}[X].$$

(2) Inner product inequality.

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

(3) Jensen inequality. For any convex function $f\colon D \to$ R, i.e., for any $x, y \in D$ and $\lambda \in (0,1)$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$, we have $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Conditional Expectations: **Key Tools of This Course**

- **We first recall the conditional probability mass function and conditional probability density function. For convenience, we restrict ourself to the simplest case $n = 2$.**
- **Assume the random vector** $(X$ + Q M 'T•B@Ð 0

- **Then the conditional expectation of X by given $Y = y_j$ is defined by**

$$\begin{aligned}\mathbb{E}[X|Y = y_j] &= \sum_i x_i q_i \\ &= \sum_i x_i \mathbb{P}(X = x_i|Y = y_j) := \boxed{h(y_j)}\end{aligned}$$

for some function h. This is reasonable since when we compute the expectation of X, its possible values are not changed, but the corresponding probabilities are changed due to the additionally information $Y = y_j$.

- **Our slogan is: conditional expectation is the expectation w.r.t. the conditional probability mass function.**

Example 15. **For some event B, define the indicator random variable $\mathbb{I}_B$ with distribution $\mathbb{P}(\mathbb{I}_B = 1) = \mathbb{P}(B) = 1 - \mathbb{P}(\mathbb{I}_B = 0)$. Prove**

$$\mathbb{E}[\mathbb{I}_B | Y = y_j] = \mathbb{P}(B | Y = y_j).$$

Proof: By definition

$$\begin{aligned} \mathbb{E}[\mathbb{I}_B | Y = y_j] &= 1 \cdot \mathbb{P}(\mathbb{I}_B = 1 | Y = y_j) + 0 \cdot \mathbb{P}(\mathbb{I}_B = 0 | Y = y_j) \\ &= \mathbb{P}(B | Y = y_j). \end{aligned}$$

□

- **Naturally, the conditional expectation of X by given Y can be uniquely defined by (Replace y_j with Y in the expression of $h(y_j)$.)**

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

Example 16 (Total Expectation Formula: Discrete Case). **Assume X, Y are two discrete random variables. Prove that**

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \sum_j \mathbb{E}[X|Y = y_j]\mathbb{P}(Y = y_j).$$

- **Why it is useful? (Assume $\mathbb{E}[X]$ exists in above formula.)**
- **We remark that if we choose $X = \mathbb{I}_B$, by above example, actually, this reduced to the total probability formula, i.e.,**

$$\mathbb{P}(B) = \sum_j \mathbb{P}(B|Y = y_j)\mathbb{P}(Y = y_j).$$

Since if we denote $A_j = \{Y = y_j\}$, then $A_j, j \geq 1$ is a division of Ω, as a consequence we immediately get the well-known form

$$\mathbb{P}(B) = \sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j).$$

Example 17.

- **We turn to the continuous case. Since the event $\{Y=y\}$ has probability zero, we can not define the conditional probability mass distribution as the discrete case. Observe that by the continuity of probability, we have (choose $\Delta y \searrow 0$ along $\mathbb{Q}$)**

$$\begin{aligned}\frac{\mathbb{P}(X\leq x, Y=y)}{\mathbb{P}(Y=y)} &= \lim_{\Delta y\searrow 0}\frac{\mathbb{P}(X\leq x, y\leq Y\leq y+\Delta y)}{\mathbb{P}(y\leq Y\leq y+\Delta y)}\\ &= \lim_{\Delta y\searrow 0}\frac{1}{\mathbb{P}(y\leq Y\leq y+\Delta y)}\int_y^{y+\Delta y}\int_{-\infty}^{x} f(u,v)\mathrm{d}u\mathrm{d}v\\ &= \frac{1}{f_Y(y)}\int_{-\infty}^{x} f(u,y)\mathrm{d}u := H(x).\end{aligned}$$

- **Obviously, $H(x)$ is a distribution function as x varies, we call it as the conditional probability distribution function of X by given $Y=y$, i.e., $X|Y=y$. Usually, we denote it as $F_{X|Y}(x,y)$.**

- **Differentiate both sides of above equation with respect to x, we get**

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(X \le x | Y = y) = \frac{f(x,y)}{f_Y(y)} := f_{X|Y}(x,y).$$

$f_{X|Y}(x,y)$ **is known as the conditional probability density function of X by given $Y = y$**

Definition 9. **The conditional expectation of X by given $Y = y$ is defined by**

$$\mathbb{E}[X|Y = y] = \int x f_{X|Y}(x,y)\mathrm{d}x := \boxed{h(y)}$$

and the conditional expectation of X by given Y is uniquely defined by $\mathbb{E}[X|Y] = \boxed{h(Y)}$.

- **Our slogan is: conditional expectation is the expectation w.r.t. the conditional probability density function.**

- Similarly as the discrete case, we also have the change of variables formula,

$$\mathbb{E}[g(X)|Y=y] = \int g(x)f_{X|Y}(x,y)\mathrm{d}x.$$

- For any event B and continuous random variable Y, how we define $\mathbb{E}[\mathbb{I}_B|Y=y]$? Intuitively (since this is true when Y is a discrete random variable), we can define

$$\mathbb{E}[\mathbb{I}_B|Y=y] = \mathbb{P}(B|Y=y).$$

- This formula can be proved by general definition of conditional expectation, we will not discuss it here. Thus for any random variable Y, we have

$$\boxed{\mathbb{E}[\mathbb{I}_B|Y=y] = \mathbb{P}(B|Y=y)}.$$

This fact will used frequently later.

Theorem 8 (Total Expectation Formula: Continuous Case). **For continuous random vector (X, Y), then**

$$\mathbb{E}[X] = \int \mathbb{E}[X|Y = y] f_Y(y) \mathrm{d}y.$$

- **Particularly, for $X = \mathbb{I}_B + Z$, where $Z \sim N(0,1)$ and is independent with Y, by the "linearity of conditional probability" (unproved fact!!!), we have**

Theorem 9 (Total Probability Formula: Continuous Case). **For any event B and continuous random variable Y, we have following total probability formula**

$$\mathbb{P}(B) = \int \mathbb{P}(B|Y = y) f_Y(y) \mathrm{d}y.$$

- **Observe that the choice of Y in total probability or expectation formula is arbitrary, so sometimes it is vital to choose suitable Y which makes the question much easier. We see some applications of total probability formula**

Example 19. **Assume X, Y are independent uniform random variables on $[0,1]$. Find the probability $\mathbb{P}(X > Y)$.**

- **A further question is, how about $\mathbb{P}(X^2 > Y)$?**

Example 20. **Let $U_i, i = 1,2,3$ are three independent uniform random variables on $[0,1]$, find $\mathbb{P}(U_1 < U_2 < U_3)$.**

Example 21. **Let $V_i, i = 1,2,3$ are three independent exponential random variables with parameters α, β and γ respectively, find $\mathbb{P}(V_1 < V_2 < V_3)$. Check you result in the special case $\alpha = \beta = \gamma$.**

More Complicated Examples (Optional)

Example 22 ($\star\star\star$). **Assume the random vector (X, Y, T) is continuous, prove that**

$$\mathbb{E}[Y|T=t] = \int \mathbb{E}[Y|T=t, X=x] f_t(x)\mathrm{d}x,$$

where f_t is the conditional density of $X|T=t$. Particularly,

$$\mathbb{P}(B|T=t) = \int \mathbb{P}(B|T=t, X=x) f_t(x)\mathrm{d}x$$

holds for any event B.

Example 23 ($\star\star\star$). **Assume $X_1, X_2 \geq 0$ are iid. random variables with common density $f(x)$. Let $U \sim Unif[0,t], t > 0$ and independents with X_1, X_2. Compute $\mathbb{P}(\{U < X_1\}|X_1 + X_2 = t)$.**

Summary (Very Important!!!)

- **In summary, in discrete case we have**

$$\mathbb{E}[g(X)] = \sum_j \boxed{\mathbb{E}[g(X)|Y = y_j]}\, \mathbb{P}(Y = y_j),$$

where

$$\boxed{\mathbb{E}[g(X)|Y = y_j]} = \sum_i g(x_i)\mathbb{P}(X = x_i|Y = y_j)$$

and particularly

$$\mathbb{P}(B) = \sum_j \mathbb{P}(B|Y = y_j)\mathbb{P}(Y = y_j).$$

Exercise 3. ☕☕ **State the corresponding similar formulas for continuous case as shown above.**

General Definition of Conditional Expectation (Optional)

- **Generally, we can define the conditional expectation of any integrable random variable X (i.e., its expectation exists) defined on $(\Omega, \mathscr{F}, \mathbb{P})$ w.r.t. some sub-σ algebra $\mathscr{G} \subset \mathscr{F}$.**

Theorem 10. **Assume X is an integrable random variable on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, $\mathscr{G} \subset \mathscr{F}$ is a sub-σ algebra of $\mathscr{F}$, then there exists an unique $\mathscr{G}$-measurable random variable η (i.e., for any Borel set A, $\eta^{-1}(A) \in \mathscr{G}$) such that for any $B \in \mathscr{G}$,**

$$\mathbb{E}[\mathbb{I}_B X] = \mathbb{E}[\mathbb{I}_B \eta].$$

- **We define η as the conditional expectation of X with respect to $\mathscr{G}$, and write as $\mathbb{E}[X|\mathscr{G}]$.**

- **Under this framework, all results discussed previously holds.** **The notation $\mathbb{E}[X|Y]$ is interpreted as $\mathbb{E}[X|\sigma(Y)]$, where**

$$\sigma(Y) := \left\{ Y^{-1}(A), A \in \mathscr{B}(\mathbf{R}) \right\}.$$

$\sigma(Y)$ is known as the σ-algebra generated by Y, which is the smallest σ algebra makes Y is measurable w.r.t. $\mathscr{F}$.

- **We can interpret $\sigma(Y)$ as the information contained in Y, and $\mathbb{E}[X|\sigma(Y)]$ means the best prediction of X based on the information Y. It can be proved that $\mathbb{E}[X|\sigma(Y)]$ must be some Borel measurable function of Y, say $h(Y)$. Of course, it is a random variable.**

- **Notation: for any set $A \in \mathscr{F}$,**

$$\mathbb{E}[X|A] = \mathbb{E}[X|\mathbb{I}_A = 1] := \mathbb{E}[X|\mathbb{I}_A]\Big|_{\mathbb{I}_A = 1}.$$

Example 24. **Assume $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space, X is an integrable valued random variable. Suppose the set $A \in \mathscr{F}$, then it is easy to check $\sigma(\mathbb{I}_A) = \sigma(A) = \{A, A^c, \emptyset, \Omega\}$ and**

$(A$

$$\mathbb{E}[X|\mathbb{I}_A] = \mathbb{E}[X|\sigma(\mathbb{I}_A)] = \mathbb{E}[X|\sigma(A)] = \frac{\mathbb{E}(X\mathbb{I}_A)}{}$$

Example 25. **How do we understand the quantity $\mathbb{E}[\mathbb{I}_B|A]$ or equivalently $\mathbb{E}[\mathbb{I}_B|\mathbb{I}_A]|_{\mathbb{I}_A=1}$? Does it reasonable we define $\mathbb{P}(B|A)$ as $\mathbb{E}[\mathbb{I}_B|A]$ and attach the meaning "if we know A happened, then the probability of B is defined as $\mathbb{P}(B|A)$"?**

Solution: **We have shown that $\mathbb{E}[\mathbb{I}_B|\mathbb{I}_A]$ is a function of $\mathbb{I}_A$ which can be used to predict $\mathbb{I}_B$ so that**

$$0 \le \mathbb{E}[\mathbb{I}_B|\mathbb{I}_A]\Big|_{\mathbb{I}_A=1} \le 1$$

and equals to 1 if $A \subset B$ and 0 if $AB = \emptyset$. These properties say $\mathbb{E}[\mathbb{I}_B|\mathbb{I}_A]$ can be used to measure (random) the uncertainty of B by using the information of A. Furthermore, we have

$$\mathbb{E}[\mathbb{I}_B|\mathbb{I}_A]\Big|_{\mathbb{I}_A=1} = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)},$$

and this definition consistents with the classical setting. In summary, such definition is plausible. □

Properties of Conditional Expectation

Theorem 11. **Following properties of conditional expectation will be used frequently. (Usually, $\mathscr{G}$ is generated by another random variable or random vector.)**

(1) **Assume** $\mathbb{E}[aX + bY|\mathscr{G}] < +\infty$ **, then**

$$\mathbb{E}[aX + bY|\mathscr{G}] = a\mathbb{E}[X|\mathscr{G}] + b\mathbb{E}[Y|\mathscr{G}].$$

(2) **Assume** $Y \in \mathscr{G}$, $\mathbb{E}[X \cdot m(Y)|\mathscr{G}] < +\infty$, **then**

$$\mathbb{E}[\mathbf{m}(\mathbf{Y})X|\mathscr{G}] = \mathbf{m}(\mathbf{Y})\mathbb{E}[X|\mathscr{G}].$$

(3) **(Total expectation formula)** $\mathbb{E}[\mathbb{E}[X|\mathscr{G}]] = \mathbb{E}[X]$.

(4) **If** X **is independent with** $\mathscr{G}$, **then** $\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X]$.

Finally Summary (Very Important!!!)

Following unified formulas should be remembered.

- **Total probability formula**

$$\mathbb{P}(B) = \int \mathbb{P}(B|X = x)\mathrm{d}F_X(x).$$

- **Total expectation formula**

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \int \mathbb{E}[Y|X = x]\mathrm{d}F_X(x).$$

-

$$\mathbb{E}[X|\mathbb{I}_A] = \mathbb{E}[X|\sigma(\mathbb{I}_A)] = \frac{\mathbb{E}[X\mathbb{I}_A]}{\mathbb{P}(A)}\mathbb{I}_A + \frac{\mathbb{E}[X\mathbb{I}_{A^c}]}{\mathbb{P}(A^c)}\mathbb{I}_{A^c}. \quad \rightsquigarrow \text{ *r.v.*}$$

and

$$\mathbb{E}[X|A] = \mathbb{E}[X|\mathbb{I}_A]\Big|_{\mathbb{I}_A=1} = \frac{\mathbb{E}[X\mathbb{I}_A]}{\mathbb{P}(A)}. \quad \rightsquigarrow \text{ *number*}$$

Convergence of Sequences of Random Variables

- **We recall three kinds of convergence modes of random variables sequence. Assume $X_n, n \geq 1$, X are random variables, F_n, F are the corresponding cumulative distribution functions.**

Definition 10. Three kinds of convergence modes.

(1) **We call X_n converges to X in distribution, if for all continuous points of F, say x, we have $F_n(x) \to F(x)$. We denote it as $X_n \to_d X$ for convenience.**

(2) **If for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$, then we say X_n converges to X in probability and write it as $X_n \to_p X$.**

(3) **We say X_n converges to X almost surely, if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$. For brevity, we write it as $X_n \to X$ a.s..**

- **We remark that for (1), we need not require $X_n, n \geq 1$, X are defined on the same probability space. Generally, we only have following relation.**

$$\boxed{X_n \to X \text{ a.s. } \Longrightarrow X_n \to_p X \Longrightarrow X_n \to_d X}.$$

Theorem 12 (Slutsky's Theorem). **Assume $X_n, Y_n, n \geq 1$ are two sequences of random variables so that $X_n \to_d X$, $Y_n \to_d c$, where c is a constant. Then**

(1) $X_n + Y_n \to_d X + c$;

(2) $X_n Y_n \to_d cX$;

(3) $X_n / Y_n \to_d X/c$ **provided** X_n / Y_n **is meaningful and** $c \neq 0$.

- **Maybe you know “ 勒 基 (Slutsky) ”.**

Definition 11 (Characteristic function). Let X be a random variable. Define its characteristic function (c.f.) as

$$\varphi(t) := \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)], \quad t \in \mathbf{R}$$

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

- c.f. always exists and $|\varphi(t)| \leq 1$ for all $t \in \mathbf{R}$.

Theorem 13. We have following important results.

(1) The characteristic function and distribution function of X are determined each other.

(2) $X_n \to_d X$ if and only if $\varphi_{X_n}(t) \to \varphi_X(t)$ for all t and $\varphi_X(t)$ is the characteristic function of some random variable and continuous in a neighborhood of zero.

Strong Law of Large Numbers: Macro Property

Theorem 14 (Kolmogorov's Strong Law of Large Numbers, SLLN).
Let $X_i, i \geq 1$ be iid. random variables with finite mean μ, then

$$(X_1 + X_2 + \cdots + X_n)/n \to \mu$$

almost surely.

- **Some other forms of law of large numbers are presented in our lecture notes.**
- **We remark that if $\mu = +\infty$, then we also have**

$$(X_1 + X_2 + \cdots + X_n)/n \to +\infty$$

almost surely.

Central Limit Theorem: Macro Property

Theorem 15 (Central Limit Theorem for iid. Case, CLT). **Assume $X_i, i \geq 1$ are iid. random variables with common mean μ and finite variance σ^2. Then**

$$\overline{S}_n := \frac{1}{\sqrt{n}\sigma}(X_1 + X_2 + \cdots + X_n - n\mu) \to_d N(0,1),$$

as $n \to \infty$. That is, for any $x \in \mathbf{R}$,

$$\lim_{n\to\infty} \mathbb{P}(\overline{S}_n \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} \mathrm{d}y := \Phi(x).$$

- **There are many ways to prove the CLT. For independent but not identically distributed case, we have the powerful Lindeberg-Feller central limit theorem, see the notes for more details.**

Interchange the Expectation and Limitation

- **Sometimes it is important to verify whether the following interchange is possible (if both exist)**

$$\boxed{\mathbb{E}\left(\lim_{n\to\infty} X_n\right) = \lim_{n\to\infty} \mathbb{E}[X_n]}. \qquad (\circledR)$$

- **($\circledR$) may not holds, for example**

Example 26. **Let $\mathbb{P}$ be the Lebesgue measure on $[0,1]$ and consider the random variables $X_n, n \geq 1$ defined as**

$$X_n(\omega) = n \cdot \mathbb{I}_{[0,n^{-1}]}(\omega)$$

then $\lim_{n\to\infty} X_n = 0$ a.s., but $\mathbb{E}[X_n] = 1$ for all n.

End of Introduction. Thanks!

A thousand journey is started by taking the first step.

千 之 ，始于 下.

–