Advanced Probability and Statistics

MAFS5020
HKUST

Kani Chen (Instructor)

PART I

PROBABILITY THEORY

Chapter 0. Review of classical probability calculations through examples.

We first review some basic concepts about probability space through examples. Then we summarize the structure of probability space and present axioms and theory.

Example 0.1. De Méré's Problem. (for the purpose of reviewing the discrete *probability space*.)

The Chevalier de Méré is a French nobleman and a gambler of the 17th century. He was puzzled by the equality of the probability of two events: At least one ace turns up with four rolls of a die, and at least one double-ace turns up in 24 rolls of two dice. His reasoning:

In one roll of a die, there is 1/6 chance of getting an ace, so in 4 rolls, there is $4 * (1/6) = 2/3$ chance of getting at least one die.

In one roll of two dice there is 1/36 chance of getting a double-ace, so in 24 rolls, there is $24*(1/36) = 2/3$ chance of getting at least one double ace.

De Méré turned to Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1665) for help. And the two mathematicians/physicists gave the right answer:

P(getting at least one ace in 4 rolls of a die) $= 1 - (1 - 1/6)^4 = 0.518$

4-rolls makes favorable bet, and 3-rolls makes not.

P(getting at least one double-ace in 24 rolls of two dice) $= 1 - (1 - 1/36)^{24} = 0.491$

25 rolls makes a favorable bet, 24 rolls make still an unfavorable bet. □

Historical remark: (Cardano's mistake) Probability theory has an infamous birth place: gambling room. The earliest publications on probability dates back to "Liber de Ludo Aleae" (the book on games of chance) by Gerolamo Cardano (1501-1576), an Italian mathematician/physician/astrologer/gambler of the time. Cardano made an important discovery of the product law of independent events, and it is believed that "Probability" was first coined and used by Cardano. Even though he made several serious mistakes that seem to be elementary nowadays, he is considered as a pioneer who first systematically computed probability of events. Pascal, Fermat, Bernoulli and de Moivre are among many other prominent developers of the probability theory. The axioms of probability was first formally and rigorously shown by Kolmogorov in the last century.

Exercise 0.1'. *Galileo's problem.* Galileo (1564-1642), the famous physicist and astronomer, was also involved in a calculation of probabilities of similar nature. Italian gamblers used to bet on the total number of spots in a roll of three dice. The question is the chance of 9 total dots the same as that of 10 total dots? There are altogether 6 combinations of total 9 dots (126, 135, 144, 234, 225, 333) and 6 combinations of total 10 dots (145, 136, 226, 235, 244, 334). This can give the false impression that the two chances are equal. Galileo gave the correct answer: 25/216 for one and 27/216 for the other. The key point here is to lay out all $6^3 = 216$ *outcomes/elements* of the *probability space* and realize that each of these 216 outcomes/elements are of the same chance 1/216. Then the chance of an *event* in the sum of the probability of the outcomes in the event.

Example 0.2. The St. Petersberg Paradox. (for the purpose of reviewing the concept of *Expectation*)

A gambler pays an entry fee M \$ to play following game: A fair coin is tossed repeated until the first head occurs and you win 2^{n-1} amount of money where n is the total number of tosses. Question: what is the "fair" amount of M?

n is a random number with $P(n = k) = 2^{-k}$ for $k = 1, 2, ...$ Therefore the "Expected Winning" is

$$E(2^{n-1}) = \sum_{k=1}^{\infty} 2^{k-1} \times \frac{1}{2^k} = \infty.$$

Notice that here I have used the expectation of a function of random variable. It appears that a "fair", but indeed naive, M should be ∞. However, by common sense, this game, despite its infinite payoff, should not worth the same as infinity.

Daniel Bernoulli (1700-1782), a Dutch-born Swiss mathematician, provided one solution in 1738. In his own words: *The determination of the value of an item must not be based on the price, but rather on the utility it yields. There is no doubt that a gain of one thousand ducats is more significant to the pauper than to a rich man though both gain the same amount.* Using a utility function, e.g., as suggested by Bernoulli himself, the logarithmic function $u(x) = \log(x)$ (known as log utility), the expected utility of the payoff (for simplicity assuming an initial wealth of zero) becomes finite:

$$E(U) = \sum_{k=1}^{\infty} u(2^{k-1}) * P(n = k) = \sum_{k=1}^{\infty} \log(2^{k-1})/2^k = \log(2) = u(2) < \infty.$$

(This particular utility function suggests that game is as useful as 2 dollars.) □

Before Bernoulli's publication in 1738, another Swiss mathematician, Gabriel Cramer, found already parts of this idea (also motivated by the St. Petersburg Paradox) in stating that *The mathematicians estimate money in proportion to its quantity, and men of good sense in proportion to the usage that they may make of it.*

Example 0.3. The dice game called "craps". (This is to review conditional probability).

Two dice are rolled repeatedly and let the total dots of the n-th roll be Z_n. If Z_1 is 2 or 3 or 12, it is an immediate loss of the game. If Z_1 is 7 or 11, it is an immediate win. Else, continue the rolls of the two dice until either Z_1 occurs, meaning a loss, or 7 occurs, meaning a win. What is the chance of a win of this game.

Solution. Write

$$\begin{aligned} P(Win) &= \sum_{k=2}^{12} P(\text{Win and } Z_1 = k) = \sum_{k=2}^{12} P(\text{Win } | Z_1 = k) P(Z_1 = k) \\ &= P(Z_1 = 7) + P(Z_1 = 11) + \sum_{k=4,5,6,8,9,10} P(\text{Win } | Z_1 = k) P(Z_1 = k) \\ &= 6/36 + 2/36 + \sum_{k=4,5,6,8,9,10} P(A_k | Z_1 = k) P(Z_1 = k) \\ &= 2/9 + \sum_{k=4,5,6,8,9,10} P(A_k) P(Z_1 = k) \end{aligned}$$

where A_k is the event that starting from the second roll, 7 dots occur before k dots. Now,

$$\begin{aligned} P(A_k) &= \sum_{j=2}^{12} P(A_k \cap \{Z_2 = j\}) = \sum_{j=2}^{12} P(A_k | \{Z_2 = j\}) P(Z_2 = j) \\ &= P(Z_2 = 7) + \sum_{\substack{j=2 \\ j \neq k, j \neq 7}}^{12} P(\text{starting from the 3rd roll, 7 occurs before k}) P(Z_2 = j) \\ &= 6/36 + P(A_k)(1 - P(Z_2 = k) - P(Z_2 = 7)). \end{aligned}$$

As a result,

$$P(A_k) = P(Z_1 = 7)/[P(Z_1 = k) + P(Z_1 = 7)].$$

And,

$$P(win) = 2/9 + \sum_{k=4,5,6,8,9,10} P(Z_1 = 7)/[P(Z_1 = k) + P(Z_1 = 7)] P(Z_1 = k) = 0.492929.$$

□

Example 0.4. Jailer's reasoning. (Bayes Probabilities)

Three men, A, B and C are in jail and one to be executed and the other two to be freed. C, being anxious, asked the jailer to tell him who of A and B would be freed. The jailer, pondering for a while, answered "for your own interest, I will not tell you, because, if I do, your chance of being executed would rise from 1/3 to 1/2." What is wrong with the jailer's reasoning?

Solution. Let AF (BF) be the event that jailer *says* A (B) to be freed. Let AE or BE or CE be the event that A or B or C to be executed. Then, $P(CE) = 1/3$. but, by the *Bayes formula* ,

$$\begin{aligned} P(CE|AF) &= \frac{P(AF|CE)P(CE)}{P(AF|AE)P(AE) + P(AF|BE)P(BE) + P(AF|CE)P(CE)} \\ &= \frac{0.5 * 1/3}{0 * 1/3 + 1 * 1/3 + 1/2 * 1/3} \\ &= 1/3 \\ &= P(CE). \end{aligned}$$

Likewise $P(CE|BF) = P(CE)$. So the "rise of probability" is false. □

Example 0.5. Buffon's needle (Continuous random variables).

Randomly drop a need of 1 cm onto a surface of many parallel straight lines that are 1 cm apart. What is the chance that the needle touches one of the lines?

Solution. Let x be the distance from the center of the needle to the nearest line. Let θ be the smaller angle of the needle with the nearest line. Then, the needle crosses a line if and only if $x/sin(\theta) \leq 0.5$.

It follows from the randomness of the drop that that θ and x are independent following the uniform distribution on $[0, \pi/2]$ and $[0, 1/2]$. Therefore,

$$P(x/sin(\theta) \leq 0.5) = \frac{4}{\pi} \int_0^{\pi/2} \int_0^{sin(\theta)/2} dx d\theta = \frac{2}{\pi} \int_0^{\pi/2} sin(\theta) d\theta = \frac{2}{\pi}.$$

The chance is $2/\pi$. □

Chapter 1. σ-algebra, measure, probability space and random variables.

This section lays the necessary rigorous foundation for probability as a mathematical theory. It begins with sets, relations among sets, measurement of sets and functions defined on the sets.

Example 1.1. (A PROTOTYPE OF PROBABILITY SPACE.) Drop a needle blindly on the interval $[0,1]$. The needle hits interval $[a,b]$, a sub-interval of $[0,1]$ with chance $b-a$. Suppose A is any subset of $[0,1]$. What's the chance or length of A?

Here, we might interpret the largest set $\Omega = [0,1]$ as the "universe". Note that not all subsets are "nice" in the sense that their volume/length can be properly assigned. So we first focus our attention on certain class of "nice" subsets.

To begin with, the "Basic" subsets are all the sub-intervals of $[0,1]$, which may be denoted as $[a,b]$, with $0 \le a \le b \le 1$. Denote $\mathcal{B}$ as the collection of all subsets of $[0,1]$, which are generated by all basic sets after $finite$ set operations. $\mathcal{B}$ is called an *algebra* of Ω.

It can be proved that any set in $\mathcal{B}$ is a finite union of disjoint intervals (closed, open or half-closed).

Still, $\mathcal{B}$ is not rich enough. For example, it does not contain the set of all rational numbers. More importantly, the limits of sets in $\mathcal{B}$ are often not in $\mathcal{B}$. This is serious restrictions of mathematical analysis.

Let $\mathcal{A}$ be the collection of all subsets of $[0,1]$, which are generated by all "basic" sets after *countably* many set operations. $\mathcal{A}$ is called Borel σ-algebra of Ω. Sets in $\mathcal{A}$ are called *Borel sets.* Limits of sets in $\mathcal{A}$ are still in $\mathcal{A}$. $(\Omega, \mathcal{A})$ is a *measurable space.*

Borel measure: any set A in $\mathcal{A}$ can be assigned a volume, denoted as $\mu(A)$, such that

(i). $\mu([a,b]) = b-a$.

(ii). $\mu(A) = \lim \mu(A_n)$ for any sequence of Borel sets $A_n \uparrow A$.

Lebesgue measure (1901): Completion of Borel σ-algebra by adding all subsets of Borel measure 0 sets, denoted as $\mathcal{F}$. Sets with measure 0 are called null sets.

Why should Borel measure or Lebesgue measure exist in general?

Caratheodory's extension theorem: extending a (σ-finite) measure on an algebra $\mathcal{B}$ to the σ-algebra $\mathcal{A} = \sigma(\mathcal{B})$.

$\Omega = [0,1]$ (the universe).

$\mathcal{B}$: an algebra (finite set operations) generated by subintervals.

$\mathcal{A}$: the Borel σ-algebra, is a σ-algebra, generated by subintervals.

$\mathcal{F}$: completion of $\mathcal{A}$, a σ-algebra, generated by $\mathcal{A}$ and null sets.

$(\Omega, \mathcal{B}, \mu)$ does not form a probability space,

$(\Omega, \mathcal{A}, \mu)$ forms a probability space.

$(\Omega, \mathcal{F}, \mu)$ forms a probability space.

Sets and set operations:

Consider Ω as the "universe", (*Beyond which is nothing.*) Write $\Omega = \{\omega\}$, ω denotes an member of the set, called *element.* Let A and B: be two subsets of Ω, called "events".

The set operations are:

intersection: $\cap$, $A \cap B$: both A and B (happens).

union: $\cup$, $A \cup B$: either A or B (happens).

complement: $A^c = \Omega \setminus A$: everything except for A, or A does not happen.

minus: $A \setminus B = A \cap B^c$: A but not B.

An elementary theorem about set operation is

DeMorgan's identity:

$$\left(\cup_{j=1}^{\infty} A_j\right)^c = \cap_{j=1}^{\infty} A_j^c, \qquad \left(\cap_{j=1}^{\infty} A_j\right)^c = \cup_{j=1}^{\infty} A_j^c.$$

In particular, $(A \cup B)^c = (A^c \cap B^c)$, i.e., $(A \cap B)^c = (A^c \cup B^c)$.

Remark. Intersection can be generated by complement and union; and union can be generated by complement and intersection.

Relation: $A \subset B$, if $\omega \in A$ ensures $\omega \in B$.

A sequence of sets $\{A_n : n \geq 1\}$ is called *increasing (decreasing)* if $A_n \subset A_{n+1}$ ($A_n \supset A_{n+1}$.)

$A = B$ if and only if $A \subset B$ and $B \subset A$.

Indicator functions. (A very useful tool to translate set operation into numerical operation)

The relation and operation of sets are equivalent to the indication set functions. For any subset $A \subset \Omega$, define its indicator function as

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{Otherwise.} \end{cases}$$

The indicator function is a function defined on Ω.

Set operations vs. function operations:

$$\begin{aligned} A \subset B &\iff 1_A \leq 1_B. \\ A \cap B &\iff 1_A \times 1_B = 1_{A\cap B} = \min(1_A, 1_B). \\ A^c = \Omega \setminus A &\iff 1 - 1_A = 1_{A^c}. \\ A \cup B &\iff 1_{A\cup B} = 1_A + 1_B, \quad \text{if } A \cap B = \emptyset \\ &\iff 1_{A\cup B} = \max(1_A, 1_B). \end{aligned}$$

Set limits.

There are two limits of sets: upper limit and low limit.

$$\begin{aligned} &\limsup A_n \equiv \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \{A_n \text{ infinitely occurs.}\} \\ &1_{\limsup A_n} = \limsup 1_{A_n} \end{aligned}$$

$\omega \in \limsup A_n$ if and only if ω belongs to infinitely many A_n.

Lower limit.

$$\begin{aligned} &\liminf A_n \equiv \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \\ = \ &\{A_n \text{ always occurs except for finite number of times.}\} \\ &1_{\liminf A_n} = \liminf 1_{A_n} \end{aligned}$$

$\omega \in \liminf A_n$ if and only if ω belongs to all but finitely many A_n.

We say the set limit of $A_1, A_2, ...$ exists if their lower limit is the same as the upper limit.

Algebra and σ-algebra

$\mathcal{A}$ is a non-empty collection (set) of subsets of Ω.

Definition. $\mathcal{A}$ is called an *algebra* if

(i). $A^c \in \mathcal{A}$ if $A \in \mathcal{A}$;

(ii). $A \cup B \in \mathcal{A}$ if $A, B \in \mathcal{A}$.

$\mathcal{A}$ is called an σ-algebra if, (ii) is strengthened as,

(iii). $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$ if $A_n \in \mathcal{A}$ for $n \geq 1$.

An algebra is closed for (finite) set operations. $\Omega \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$.

A σ-algebra is closed for *countable* operations.

$(\Omega, \mathcal{A})$ is called a measurable space, if $\mathcal{A}$ is a σ-algebra of Ω.

Measure, measure space and probability space.

$\mathcal{A}$, containing $\emptyset$, is a non-empty collection (set) of subsets of Ω. μ is a nonnegative set function on $\mathcal{A}$.

μ is called a *measure*, if

(i). $\mu(\emptyset) = 0$.

(ii). $\mu(A) = \sum_{n=1}^{\infty} \mu(A_n)$ if $A, A_1, A_2, ...$ are all in $\mathcal{A}$ and $A_1, A_2, ...$ are disjoint.

$(\Omega, \mathcal{A}, \mu)$ is called a measure space, if μ is a measure on $\mathcal{A}$ and $\mathcal{A}$ is a σ-algebra of Ω.

$(\Omega, \mathcal{A}, P)$ is called a probability space if $(\Omega, \mathcal{A}, P)$ is a measure space and $P(\Omega) = 1$.

For probability space $(\Omega, \mathcal{A}, P)$, Ω is called sample space, every A in $\mathcal{A}$ is an event, and $P(A)$ is the probability of the event, the chance that it happens.

Random variable (r.v.).

Loosely speaking, given a probability space $(\Omega, \mathcal{F}, P)$, a random variable (r.v.) X is defined as a real-valued function of Ω, satisfying certain measurability condition. Loosely speaking, viewing $X = X(\omega)$ as a mapping from Ω to R, the real line, then $X^{-1}(B)$ must be in $\mathcal{F}$ for all Borel sets B. (Borel sets on real line are the σ-algebra generated by intervals, i.e., the sets generated by countable operations on intervals).

A random variable X defined on a probability space $(\Omega, \mathcal{A}, P)$ is a function defined on Ω, such that $X^{-1}(B) \in \mathcal{A}$ for every interval B on $[-\infty, \infty]$, where $X^{-1}(B) = \{\omega : X(\omega) \in B\}$. (*We need to identify its probability.*)

$X^{-1}(B)$ is called the inverse image of B.

$X = X(\cdot)$ can be viewed as a map or transformation from $(\Omega, \mathcal{A})$ to $(R, \mathcal{B})$, where $R = [-\infty, \infty]$ and $\mathcal{B}$ is the σ-algebra generated by the intervals in R.

X is a *measurable map/transformation* since $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$ (DIY.)

Because $\mathcal{A}$ is a σ-algebra, the upper and lower limits of X_n is a r.v. if X_n are r.v.s., and the algebraic operations: $+, -, \times, /$, of r.v.s are still r.v.s.

Measurable map and random vectors.

$f(\cdot)$ is called a *measurable map/transformation/function* from a measurable space $(\Omega, \mathcal{A})$ to another measurable space $(S, \mathcal{S})$, if $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{S}$. i.e. $\{w : f(w) \in B\} \in \mathcal{A}$.

X is called a random vector of p dimension if it is a measurable map from a probability space $(\Omega, \mathcal{A}, P)$ to $(R^p, \mathcal{B}^p)$, where $\mathcal{B}^p$ is the Borel σ-algebra in p dimensional real space, $R^p = [-\infty, \infty]^p$.

Proposition 1.1 If $X = (X_1, ..., X_p)$ is a random vector of p dimension on a probability space $(\Omega, \mathcal{A}, P)$, and $f(\cdot)$ is measurable function from $(R^p, \mathcal{B}^p)$ to $(R, \mathcal{B})$, then $f(X)$ is a random variable.

Proof. For any Borel set $B \in \mathcal{B}$,

$$\{\omega : f(X(\omega)) \in B\} = \{\omega : X(\omega) \in f^{-1}(B)\} \in \mathcal{A}$$

since $f^{-1}(B) \in \mathcal{B}^p$. □

Proposition 1.2 If $X_1, X_2, ...$ are r.v.s. So are

$$\inf_n X_n, \quad \sup_n X_n \quad \liminf_n X_n \quad \text{and} \quad \limsup_n X_n.$$

Proof. Let the probability space be $(\Omega, \mathcal{A}, P)$. For any x,

$$\{\omega : \inf_n X_n(\omega) \geq x\} = \cap_n \{\omega : X_n(\omega) \geq x\} \in \mathcal{A};$$

$$\{\omega : \sup_n X_n(\omega) \leq x\} = \cap_n \{\omega : X_n(\omega) \leq x\} \in \mathcal{A};$$

$$\{\liminf_n X_n > x\} = \cup_n \{\inf_{k \geq n} X_k > x\} \in \mathcal{A};$$

$$\{\limsup_n X_n < x\} = \cup_n \{\sup_{k \geq n} X_k < x\} \in \mathcal{A}.$$

Therefore, $\inf_n X_n, \sup_n X_n, \liminf_n X_n$ and $\limsup_n X_n$ are r.v.s. □

Proposition 1.3 Suppose X is a map from a measurable space $(\Omega, \mathcal{A})$ to another measurable space $(\mathbf{S}, \mathcal{S})$. If $X^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{C}$ and $\mathcal{S} = \sigma(\mathcal{C})$. Then, X is a measurable map, i.e., $X^{-1}(S) \in \mathcal{A}$ for every $S \in \mathcal{S}$. In particular, when $(\mathbf{S}, \mathcal{S}) = ([-\infty, \infty], \mathcal{B})$, $X^{-1}([-\infty, x]) \in \mathcal{A}$ for every x is enough to ensure X is a r.v..

Proof. Note that $\sigma(\mathcal{C})$, the σ-algebra generated by $\mathcal{C}$, is defined mathematically as the smallest σ-algebra containing $\mathcal{C}$.

Set $\mathcal{B}^* = \{B \in \mathcal{S} : X^{-1}(B) \in \mathcal{A}\}$.

We first show $\mathcal{B}^*$ is a σ-algebra. Observe that

(i). for any $B \in \mathcal{B}^*$, $X^{-1}(B) \in \mathcal{A}$ and, therefore, $X^{-1}(B^c) = (X^{-1}(B))^c \in \mathcal{A}$;

(ii). for any $B_n \in \mathcal{B}^*$, $X^{-1}(B_n) \in \mathcal{A}$ and $X^{-1}(\cup_n B_n) = \cup_n X^{-1}(B_n) \in \mathcal{A}$.

Consequently, $\mathcal{B}^*$ is a σ-algebra. Since $\mathcal{C} \subset \mathcal{B}^* \subset \mathcal{S}$, it follows that $\mathcal{B}^* = \mathcal{S}$. □

set operations.

a σ-algebra of Ω and P a set function such that

DIY Exercises:

Exercise 1.1 ⋆⋆ Show $1_{\liminf A_n} = \liminf 1_{A_n}$ and DeMorgen's identity.

Exercise 1.2 ⋆⋆ Show that, the so called "countable additivity" or "σ-additivity", ($P(\cup_n A_n) = \sum_n P(A_n)$ for countable disjoint $A_n \in \mathcal{A}$), is equivalent to "finite additivity" plus "continuity" (if $A_n \downarrow \emptyset$, then $P(A_n) \to 0$.)

Exercise 1.3 ⋆⋆⋆ *(Completion of a Probability space)* Let $(\Omega, \mathcal{F}, P)$ be a probability space. Define

$$\bar{\mathcal{F}} = \{A : P(A \setminus B) + P(B \setminus A) = 0, \text{ for some} B \in \mathcal{F}\},$$

And for each $A \in \bar{\mathcal{F}}$, $P(A)$ is defined as $P(B)$ for the B given above. Prove that $(\Omega, \bar{\mathcal{F}}, P)$ is also a probability space. (Hint: need to show that $\mathcal{F}$ is a σ-algebra and that P is a probability measure.)

Exercise 1.4 ⋆⋆⋆ If X_1 and X_2 are two r.v.s, so is $X_1 + X_2$. (Hint: cite Propositions 1.1 and 1.3)

Chapter 2. Distribution, expectation and inequalities.

Expectation, also called mean, of a random variable is often referred to as the location or center of the random variable or its distribution. To avoid some non-essential trivialities, unless otherwise stated, the random variables will usually be assumed to take finite values and those taking values $-\infty$ and ∞ are considered as *r.v.*s in extended sense.

(i). *Distribution.*

Recall that, given a probability space $(\Omega, \mathcal{F}, P)$, a random variable (*r.v.*) X is defined as a real-valued function of Ω, satisfying certain measurability condition. The *cumulative distribution function* of X is then

$$F(t) = P(X \le t) = P(\{w \in \Omega : X(w) \le t\}) = P(X^{-1}((-\infty, t])), \quad t \in (-\infty, \infty).$$

$F(\cdot)$ is then a right-continuous function defined on the real line $(-\infty, \infty)$.

REMARK. The distribution function of a single r.v. may be considered as complete profile/description of the r.v.. The distribution function $F(\cdot)$ defines a probability measure on $(-\infty, \infty)$. This is the *induced measure*, induced by the random variable as a map/function from the probability measure P on $(\Omega, \mathcal{F}, P)$ to $((-\infty, \infty), \mathcal{B}, F)$. In this sense, the original probability space is often left unspecified or seemingly irrelevant when dealing with one single random variable.

We often call a random variablediscrete random variableif it takes countable number of values, and call a random variablecontinuous random variableif the chance it takes any particular value is 0. In statistics, continuous random variableis often, by default, given a density function. In general, continuous random variablemay not have a density function (with respect to Legbegue measure). An example is the Cantor measure.

For two random variables X and Y, their joint c.d.f. is

$$F_{X,Y}(t, s) = P(X \le t \,\text{and}\, Y \le s) = P(X^{-1}((-\infty, t]) \cap Y^{-1}((-\infty, s])), \; t, s \in (-\infty, \infty).$$

Joint c.d.f can be extended for finite number of variables in a straightforward fashion. If the (joint) c.d.f. is differentiable, the derivative is then called (joint) density.

(ii). *Expectation.*

Definitions. For a nonnegative random variable$X$ with c.d.f F, its expectation is defined as

$$E(X) \equiv \int_0^\infty x dF(x).$$

In general, let $X^+ = X1_{\{X \ge 0\}}$, $X^- = -X1_{\{X \le 0\}}$,

$$E(X) \equiv E(X^+) - E(X^-).$$

If $E(X^+) = \infty = E(X^-)$, $E(X)$ does not exist.

A more original definition of the expectation is through that of Lebesgue integral: for nonnegative X,

$$\begin{aligned} E(X) &\equiv \int X(w) dP(w) \qquad \text{formally} \\ &\equiv \lim_{m \to \infty} \sum_{k=0}^{\infty} \frac{k}{2^m} P\Big(\frac{k}{2^m} < X \le \frac{n+1}{2^m}\Big). \end{aligned}$$

If X takes ∞ with positive probability, $E(X^+) = \infty$. Note that X has finite mean is equivalent to $E|X| < \infty$. And the mean of X does not exist is the same as $E(X^+) = E(X^-) = \infty$.

The expectation defined above is mathematically an integral or summation with respect to certain probability measure induced by the random variable. In layman's words, it is the weighted "average" of the values taken by the *r.v.*, weighted by chances which sum up to 1.

Some basic properties of expectation:

(1). $E(f(X)) = \int f(x)dF(x)$ where F is the c.d.f. of X.

(2). If $P(X \leq Y) = 1$, then $E(X) \leq E(Y)$. If $P(X = Y) = 1$ then $E(X) = E(Y)$.

(3). $E(X)$ is finite if and only if $E(|X|)$ is finite.

(4). (Linearity) $E(aX + bY) = aE(X) + bE(Y)$.

(5). If $a \leq X \leq b$, then $a \leq E(X) \leq b$.

(iii). *Some typical distributions of random variables.*

(1.) Commonly used discrete distributions:

Bernoulli: $X \sim Bin(1, p)$. $P(X = 1) = p = 1 - P(X = 0)$. $E(X) = p$ and $\text{var}(X) = p(1 - p)$.

Binomial: $X \sim Bin(n, p)$. $X = \sum_{i=1}^{n} x_i$ and x_i are iid with $B(1, p)$ (the number of successes of n Bernoulli trials.

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

$E(X) = np$. $\text{var}(X) = np(1 - p)$.

Poisson: $X \sim \mathcal{P}(\lambda)$. $E(X) = \text{var}(X) = \lambda$.

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

Key fact: $B(n, p) \to \mathcal{P}(\lambda)$ if $n \to \infty$, $np \to \lambda$. (Law of rare events.)

Geometric: $X \sim G(p)$: time to the first success in a series of Bernoulli trials.

$$P(X = k) = (1 - p)^{k-1} p, \qquad k = 1, 2, ...$$

$E(X) = 1/p, \text{var}(X) = (1 - p)/p^2$.

Negative binomial: $X \sim NB(p, r)$: time to the first r successes in a series of Bernoulli trials. Therefore $X = \sum_{j=1}^{r} \xi_j$ where ξ_j are iid $\sim G(p)$.

$$P(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}, \qquad k = r, r + 1, ...$$

$E(X) = r/p$ and $\text{var}(X) = r(1 - p)/p^2$.

Hyper-geometric: $X \sim HG(r, n, m)$: the number of black balls when r balls are taken without replacement from an urn containing n black balls and m white balls.

$$P(X = k) = \binom{n}{k}\binom{m}{r-k} / \binom{n+m}{r}, \qquad k = 0 \vee (r - m), 1, ..., r \wedge n.$$

$E(X) = rn/(m + n)$ and $\text{var}(X) = rnm(n + m - r)/[(n + m)^2(n + m - 1)]$.

(2) Commonly used continuous distributions:

Uniform: $X \sim Unif[a, b]$

$$f(x) = (b - a)1_{\{x \in [a,b]\}}$$

$E(X) = (a+b)/2$ and $\text{var}(X) = (b-a)^2/12$.

Normal: $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$ and $\text{var}(X) = \sigma^2$. Central limit theorem.

$$f(x) = 1/\sqrt{2\pi\sigma^2} e^{-(x-\mu)^2/(2\sigma^2)}, \qquad x \in (-\infty, \infty).$$

Exponential: $X \sim \mathcal{E}(\lambda)$. Density:

$$f(x) = e^{-x/\lambda}/\lambda, \qquad x > 0$$

$E(X) = \lambda$ and $\text{var}(X) = \lambda^2$. No memory: $(X-t) \,|\, \{X \geq t\} \sim \mathcal{E}(\lambda)$.

Gamma: $\Gamma(\alpha, \gamma)$. Density:

$$f(x) = \frac{1}{\Gamma(\alpha)\gamma} x^{\alpha-1} e^{-x/\gamma}, \qquad x > 0.$$

$\mathcal{E}(\lambda) = \Gamma(1, \lambda)$, $\chi^2_n = \Gamma(n/2, 2)$. Sum of independent $\Gamma(\alpha_i, \gamma)$ follows $\Gamma(\sum_i \alpha_i, \gamma)$.

Beta: $B(\alpha, \beta)$. Density:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad x \in [0,1]$$

$\xi/(\xi+\eta) \sim B(\alpha, \beta)$ where $\xi \sim \Gamma(\alpha, \gamma)$ and $\eta \sim \Gamma(\beta, \gamma)$ are independent. $X_{(k)} \sim B(k-1, n-k+1)$ as the k-th smallest of $X_1, ..., X_n$ iid $\sim Unif[0,1]$

Cauchy: density $f(x) = 1/[\pi(1+x^2)]$. Symmetric about 0, but expectation and variance not exist.

χ^2_n (with d.f. n): sum of n i.i.d standard normal r.v.s. χ^2_2 is $\mathcal{E}(2)$.

t_n (with d.f n): $\xi/\sqrt{\eta/n}$ where $\xi \sim N(0,1)$, $\eta \sim \chi^2_n$ and ξ and η are independent.

$F_{m,n}$ (with d.f. (m,n)): $(\xi/m)/(\eta/n)$ where $\xi \sim \chi^2_m$, $\eta \sim \chi^2_n$ and ξ and η are independent.

(iv). Some basic inequalities:

Inequalities are extremely useful tools in theoretical development of probability theory. For simplicity of notation, we use $\|X\|_p$, which is also called L_p norm if $p \geq 1$, to denote $[E(|X|^p)]^{1/p}$ for a r.v. X. In what follows, X and Y are two random variables.

(1) *the Jensen inequality*: Suppose $\psi(\cdot)$ is a convex function and X and $\psi(X)$ have finite expectation. Then $\psi(E(X)) \leq E(\psi(X))$.

Proof. Convexity implies for every a, there exists a constant c such that $\psi(x) - \psi(a) \geq c(x-a)$. Let $a = E(X)$ and $x = X$, the right hand side is mean 0. So Jensen's inequality follows. □

(2). *the Markov inequality*: For any $a > 0$, $P(|X| \geq a) \leq 1/aE(|X|)$.

Proof. $aP(|X| \geq a) = E(a1_{\{|X| \geq a\}} \leq E(|X|1_{\{|X| \geq a\}}) \leq E(|X|)$. □

(3). *the Chebyshev (Tchebychev) inequality*: for $a > 0$,

$$P(|X - E(X)| \geq a) \leq \text{var}(X)/a^2$$

Proof. The inequality holds if $\text{var}(X) = \infty$. Assume $\text{var}(X) < \infty$, then $E(X)$ is finite and $Y \equiv (X - E(X))^2$ is well defined. It follows from the Markov inequality that

$$P(|X - E(X)| \geq a) = P(Y \geq a^2) \leq E(Y)/a^2 = \text{var}(X)/a^2.$$

□

(4). *the Hölder inequality*: for $1/p + 1/q = 1$ with $p > 0$ and $q > 0$,

$$E|XY| \le \|X\|_p \|Y\|_q$$

Proof. Observe that for any two nonnegative numbers a and b, $ab \le a^p/p + b^q/q$. (This is a result of the concavity of the log-function. please DIY.) Let $a = |X|/\|X\|_p$ and $b = |Y|/\|Y\|_q$ and take expectation on both sides. The Hölder inequality follows. □

(5). *the Schwarz inequality*:

$$E(|XY|) \le [E(X^2)E(Y^2)]^{1/2}.$$

Proof. A special case of the Hölder inequality. □

(6). *the Minkowski inequality*: for $p \ge 1$,

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_p.$$

Proof. If $p = 1$, the inequality is trivial. Assume $p > 1$. Let $q = p/(p-1)$. Then $1/p + 1/q = 1$. By the Hölder inequality,

$$E[|X||X+Y|^{p-1}] \le \|X\|_p \||X+Y|^{p-1}\|_q = \|X\|_p \{E[|X+Y|^{(p-1)q}]\}^{1/q} = \|X\|_p \{E[|X+Y|^p]\}^{(p-1)/p}.$$

Likewise,

$$E[|Y||X+Y|^{p-1}] \le \|Y\|_p \{E[|X+Y|^p]\}^{(p-1)/p}.$$

Summing up the above two inequalities leas to

$$E(|X+Y|^p) \le (\|X\|_p + \|Y\|_p)\{E[|X+Y|^p]\}^{(p-1)/p},$$

and the Minkowski inequality follows. □

Remark. Jensen's inequality is a powerful tool. For example, straightforward applications include

$$[E(|X|)]^p \le E(|X|^p), \quad \text{for } p \ge 1,$$

which implies

$$\|X\|_p \le \|Y\|_q, \quad \text{for } 0 < p < q.$$

Moreover,

$$E(\log(|X|)) \le \log(E(|X|)).$$

If $E(X)$ exists,

$$E(e^X) \ge e^{E(X)}.$$

These inequalities are all very commonly used. For example, the validity of the maximum likelihood likelihood estimation essentially rests on the fact,

$$E\log\Big(\frac{f_\theta(X)}{f_{\theta_0}(X)}\Big) \le \log E\Big(\frac{f_\theta(X)}{f_{\theta_0}(X)}\Big) = \log\Big(\int \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)dx\Big) = \log\Big(\int f_\theta(x)dx\Big) = \log(1) = 0,$$

which is a result of Jensen's inequality. Here $f_\theta(\cdot)$ is a parametric family of density of X with θ_0 being the true value of θ.

The Markov inequality, despite its simplicity, shall be frequently used in the order of a sequence of random variables, especially when coupled with the technique of truncation. The Chebyshev inequality is so mighty that, as an example, it directly proves the weak law of large numbers.

The Schwarz inequality shows that covariance is an inner product, and, furthermore, the space of mean 0 *r.v.*s with finite variances forms a Hilbert space. The Minkowsky inequality is the triangle inequality for L_p norm, without which L_p cannot be a norm.

DIY Exercises.

Exercise 2.1. ⋆ Suppose X is a r.v. taking values on all rational numbers on $[0,1]$, Specifically, $P(X = q_i) = p_i > 0$ where $q_1, q_2, ...$ denotes all rational numbers on $[0,1]$. Then, the c.d.f of X is continuous at irrational numbers and discontinuous at rational numbers.

Exercise 2.2. ⋆⋆⋆ Show $\text{var}(X^+) \leq \text{var}(X)$ and $\text{var}(\min(X, c)) \leq \text{var}(X)$ where c is any constant.

Exercise 2.3. ⋆⋆⋆ (*Generalizing Jensen's inequality*). Suppose $g(\cdot)$ is a convex function and X is a random variable with finite mean. Then, for any constant c,

$$Eg(X - E(X) + c) \geq g(c).$$

Exercise 2.4. ⋆⋆⋆ *Lyapunov (Liapounov)* : Show that the function $\log E(|X|^p)$ is a convex function of p on $[0, \infty)$. Or, equivalently, for any $0 < s < m < l$, show

$$E(|X|^m) \leq [E(|X|^s)]^r [E(|X|^l)]^{1-r}$$

where $r = (l - m)/(l - s)$. (Hint: use the H‡older inequality on

$$E(|X|^{\lambda p_1 + (1-\lambda)p_2}) \leq [E(|X|^{p_1})]^{\lambda} [E(|X|^{p_2})]^{1-\lambda}$$

for positive p_1, p_2 and $0 < \lambda < 1$.)

Chapter 3. Convergence

Unlike convergence of a sequence of numbers, the convergence of a sequence of r.v.s at least has four commonly used modes: almost sure convergence, in probability convergence, L_p convergence and in distribution convergence. The first is sometimes called convergence almost everywhere or almost certain and the last convergence in law.

(i). Definitions

In what follows, we give definitions. Suppose $X_1, X_2, ...$ are a sequence of r.v.s.

$X_n \to X$ almost surely, (*a.s.*) if $P(\{\omega : X_n(\omega) \to X(\omega)\}) = P(X_n \to X) = 1$. Namely, *a.s.* convergence is a point-wise convergence "everywhere" except for a null set.

$X_n \to X$ in probability, if $P(|X_n - X| > \epsilon) \to 0$ for any $\epsilon > 0$.

$X_n \to X$ in L_p, if $E(|X_n - X|^p) \to 0$.

$X_n \to X$ in distribution. There are four equivalent definitions:

1). For every continuity point t of F, $F_n(t) \to F(t)$, where F_n and F are c.d.f of X_n and X.

2). For every closed set B, $\limsup_n P(X_n \in B) \leq P(X \in B)$.

3). For every open set B, $\liminf_n P(X_n \in B) \geq P(X \in B)$.

4). For every continuous bounded function $g(\cdot)$, $E(g(X_n)) \to E(g(X))$.

Remark. The L_p convergence preclude the limit X taking values of infinity with positive chances. Sometimes in some textbooks, a sequence of numbers going to infinity is called convergence to infinity rather than divergence to infinity. If this is the case, the limit X can be ∞ or $-\infty$, for a.s. convergence and, by slightly modifying the definition, for in probability convergence. For example, $X_n \to \infty$ in probability is naturally defined as, for any $M > 0$, $P(X_n > M) \to 1$. Convergence in distribution only has to do with distributions.

(ii). Convergence theorems.

The following three theorems/lemma, tantamount to their analogues in real analysis, play important role in the technical development of probability theory.

(1). *Monotone convergence theorem.* If $X_n \geq 0$, and $X_n \uparrow X$, then $E(X_n) \uparrow E(X)$.

Proof. $E(X_n) \leq E(X)$. For any $a < E(X)$, there exists a N and m such that $\sum_{i=0}^{N} \frac{i}{2^m} P\Big(\frac{i}{2^m} < X(w) \leq \frac{i+1}{2^m}\Big) > a$. But $P\Big(\frac{i}{2^m} < X_n(w) \leq \frac{i+1}{2^m}\Big) \to P\Big(\frac{i}{2^m} < X(w) \leq \frac{i+1}{2^m}\Big)$ (why?). Therefore, $\lim E(X_n) \geq a$. Hence, $E(X_n) \to E(X)$. □

(2). *Fatou's lemma.* If $X_n \geq 0$, a.s., then

$$E(\liminf X_n) \leq \liminf E(X_n)$$

Proof. Let $X_n^* = \inf(X_k : k \geq n)$, then $X_n^* \uparrow \liminf X_n$, so the Monotone convergence theorem, $E(X_n^*) \uparrow E(\liminf X_n)$. On the other hand, $X_n^* \leq X_n$ so, $E(X_n^*) \leq E(X_n)$. As a result, $E(\liminf X_n) \leq \liminf E(X_n)$. □

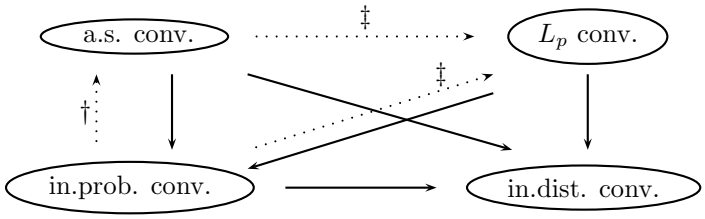
(3). *Dominated convergence theorem.* If $|X_n| \leq Y$, $E(Y) < \infty$, and $X_n \to X$ a.s., then $E(X_n) \to E(X)$.

Proof. Observe that $Y - X_n \geq 0 \leq Y + X_n$. By Fatou's lemma, $E(Y - \lim X_n) \leq \liminf E(Y - X_n)$, leading to $E(X) \geq \limsup E(X_n)$. Likewise $E(Y + \lim X_n) \leq \liminf E(Y + X_n)$, leading to $E(X) \leq \liminf E(X_n)$. Consequently, $E(X_n) \to E(X)$. □

The essence of the above convergence theorems is to use a bound, upper or lower, to ensure the desired convergence in expectation. These bounds, lower bounds as 0 in the monotone convergence theorem and the Fatou lemma, and both lower and upper bounds in the dominated convergence theorem, can actually be relaxed; see DIY exercises. The most general extension is through the concept of uniform integral r.v.s, which shall be introduced later if necessary.

(iii). Relations between convergence modes.

The relations are partly illustrated in the following diagram:



†: exist a subsequence that converges a.s. ‡: if $|X_n| \leq Y$ where $Y \in L_p$.

(iv) Some examples.

We use following examples to clarify the above diagram.

a). in prob. conv. but not a.e. conv.

Let $\xi \sim Unif[0,1]$. Set $X_{2^j+k} = 1$ if $\xi \in [k/2^j, (k+1)/2^j]$ and 0 otherwise, for all $0 \leq k \leq 2^j - 1$ and $j = 0, 1, 2, ...$ Then, $X_n \to 0$ in probability as $n \to \infty$, but $X_n \nrightarrow 0$, a.e.. In fact, $P(X_n \to 0) = 0$.

Let ξ_n be i.i.d $\sim Unif[0,1]$. Let $X_n = 1$ if $\xi_n \leq 1/n$ and 0 otherwise. Then $X_n \to 0$ in probability, but $X_n \nrightarrow 0$, a.e. by Borel-Contelli lemma.

b). in distribution conv. but not in probability conv..

This is in fact quite trivial. Any sequence of (non-constant) i.i.d. random variables converge in distribution, but not in probability. Observe that convergence in distribution only concerns the distribution. The variables even do not have to be in the same probability space.

c). a.s. but not L^p conv.

Let $\xi \sim Unif[0,1]$. Let $X_n = e^n$ if $\xi \leq 1/n$ and 0 otherwise. Then $X_n \to 0$ a.s. but $E(|X_n|^p|) = e^{np}/n \to \infty$.

(v). Technical proofs.

①. a.s. convergence $\Longrightarrow$ in probability convergence.

Proof. Let $A_n = \{|X_n - X| > \epsilon\}$. a.s. convergence implies $P(A_n, i.o.) = 0$. But $\{A_n, i.o.\} = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$. So $0 = P(A_n, i.o.) = \lim_n P(\cup_{k=n}^{\infty} A_k) \geq \limsup_n P(A_n)$.

②. L^p convergence $\Longrightarrow$ in prob convergence.

Proof. $0 \leftarrow E(|X_n - X|^p) \geq E(|X_n - X|^p 1_{\{|X_n - X| > \epsilon\}}) \geq \epsilon^p P(|X_n - X| > \epsilon)$.

③. in prob convergence $\Longrightarrow$ in distribution convergence.

Proof. For any t, and any $\epsilon > 0$, $\limsup P(X_n \leq t) \leq \limsup P(\{X_n \leq t\} \cap \{X \leq X_n + \epsilon\}) \leq P(X \leq t + \epsilon)$. Let $\epsilon \downarrow 0$, we have $\limsup P(X_n \leq t) \leq P(X \leq t)$. Likewise $\limsup P(-X_n \leq -t) \leq P(-X \leq -t)$. (Why?) Then $\liminf P(X_n < t) \geq P(X < t)$. Suppose now, t is a continuity point of X. Then $P(X < t) = P(X \leq t)$. As a result, $\lim_n P(X_n \leq t) = P(X \leq t)$.

④. in prob convergence $\Longrightarrow$ existence of a subsequence that converges a.s.

Proof. Let $\epsilon_k \downarrow 0$. Since $P(|X_n - X| > \epsilon_k) \to 0$ as $n \to \infty$, there exists an n_k such that $P(|X_{n_k} - X| > \epsilon_k) < 2^{-k}$. Therefore $\sum_{k=1}^{\infty} P(|X_{n_k} - X| > \epsilon_k) < \infty$, which implies by the Borel-Contelli lemma, which is introduced in the next section, that $P(|X_{n_k} - X| > \epsilon_k, i.o.) = 0$. This means that, with probability 1, $|X_{n_k} - X| \leq \epsilon_k$ for all large k. This is tantamount to $X_{n_k} \to X$ a.s..

⑤. L^p convergence $\Longrightarrow$ L^q convergence for $p > q > 0$.

Proof. Let $Y_n = |X_n - X|$. For any $\epsilon > 0$, $E(Y_n^q) \leq \epsilon + E(Y_n^q 1_{\{Y_n \geq \epsilon\}}) \leq \epsilon + E(Y_n^q 1_{\{Y_n \geq 1\}} + P(\epsilon \leq Y_n \leq 1) \leq \epsilon + E(Y_n^p 1_{\{Y_n \geq 1\}} + P(\epsilon \leq Y_n) \to \epsilon$ as $n \to \infty$. Since $\epsilon > 0$ is arbitrary, it follows that $X_n \to X$ in L^q.

⑥. Suppose $|X_n| \leq c > 0$ a.s., then, in probability convergence $\Longleftrightarrow$ L^p convergence for all (any) $p > 0$.

Proof. $\Longleftarrow$ follows from ②. And $\Longrightarrow$ follows from the dominated convergence theorem.

⑦. The four equivalent definitions of in distribution convergence.

Proof. 2) $\Longleftrightarrow$ 3). The complement of any closed set is open. Likewise, the complement of any closed set is open.

1) $\Longrightarrow$ 3). Continuity points of F are dense (why?). Consider interval $(-\infty, t)$, there exists continuity points $t_k \uparrow t$. Then,

$$\liminf_n P(X_n \in (-\infty, t)) \geq \liminf_n P(X_n \in (-\infty, t_k]) = P(X \in (-\infty, t_k]) \to P(X \in (-\infty, t)).$$

The result can be extended for general open sets. We omit the proof.

3) $\Longrightarrow$ 1). Suppose t is a continuity point. Then $\limsup_n F_n(t) \leq F(t)$ by 2) and the equivalency of 2) and 3). $\liminf_n F_n(t) \geq \liminf_n P(X_n < t) \geq P(X < t) = F(t)$ as t is a continuity point. So 1) follows.

4) $\Longrightarrow$ 1). Let t be a continuity point of F. For any small $\epsilon > 0$, choose a non-increasing continuous function f of x which is 1 for $x < t$, and is 0 for $x > t + \epsilon$. Then, $P(X_n \leq t) \leq E(f(X_n)) \to E(f(X)) \leq P(X \leq t + \epsilon)$. Therefore the $\limsup P(X_n \leq t) \leq P(X \leq t)$. Likewise (how?), one can show $\liminf P(X_n \leq t) \geq P(X \leq t)$. The desired convergence follows.

1) $\Longrightarrow$ 4). Continuity points of the cdf of X are dense (why?). Suppose $|f(t)| < c$. Choose continuity points $-\infty = t_0 < t_1, ... < t_K < t_{K+1} = \infty$ such that $F(t_1) < \epsilon > 1 - F(t_K)$, and $|f(t) - f(s)| < \epsilon$ for any $t, s \in [t_j, t_{j+1}]$ for $j = 1, ..., K-1$. Then,

$$\begin{aligned}
& |E(f(X_n)) - E(f(X))| = |\int f(t) dF_n(t) - \int f(t) dF(t)| \\
\leq & \sum_{j=0}^{K} |\int_{t_j}^{t_{j+1}} f(t)[dF_n(t) - dF(t)]| \\
\leq & 2c\epsilon + \sum_{j=1}^{K-1} |\int_{t_j}^{t_{j+1}} f(t)[dF_n(t) - dF(t)]| \\
\leq & 2c\epsilon + + \sum_{j=1}^{K-1} |\int_{t_j}^{t_{j+1}} f(t_j)[dF_n(t) - dF(t)]| \\
& + \sum_{j=1}^{K-1} |\int_{t_j}^{t_{j+1}} [f(t) - f(t_j)][dF_n(t) - dF(t)]|
\end{aligned}$$

$$
\begin{aligned}
&\leq 2c\epsilon + \sum_{j=1}^{K-1} c\Big|\int_{t_j}^{t_{j+1}} [dF_n(t) - dF(t)]\Big| + \sum_{j=1}^{K-1} \epsilon \int_{t_j}^{t_{j+1}} [dF_n(t) + dF(t)] \\
&\to 2c\epsilon + 2\epsilon \int_{t_1}^{t_K} dF(t) \qquad \text{as } n \to \infty. \\
&\leq (2c+1)\epsilon,
\end{aligned}
$$

which can be arbitrarily small.

DIY Exercises.

Exercise 3.1 ⋆⋆ Suppose $X_n \geq \eta$, with $E(\eta^-) < \infty$. Show $E(\liminf X_n) \leq \liminf E(X_n)$.

Exercise 3.2 ⋆⋆ Show the dominated convergence theorem still holds if $X_n \to X$ in probability or in distribution.

Exercise 3.3 ⋆⋆⋆ Let $S_n = \sum_{i=1}^n X_i$. Raise a counter-example to show $S_n/n \not\to 0$ in probability but $X_n \to 0$ in probability.

Exercise 3.4 ⋆⋆⋆ Let $S_n = \sum_{i=1}^n X_i$. Show that $S_n/n \to 0$ *a.s.* if $X_n \to 0$ *a.s.*, and $S_n/n \to 0$ in L_p if $X_n \to 0$ in L_p for $p \geq 1$.

Chapter 4. Independence, conditional expectation, Borel-Cantelli lemma and Kolmogorov 0-1 laws.

(i). Conditional probability and independence of events.

For any two events, say A and B, the conditional probability of A given B is defined as

$$P(A|B) = P(A \cap B)/P(B), \text{ if } P(B) \neq 0.$$

This is the chance of A to happen, given B has happened.

In common sense, the independence between events A and B should be, information about event B happens/or not, does not change the chance of A to happen/or not, and vice versus. In other words, whether B (A) happens or not does not contain any information about whether A (B) happens. Therefore the definition of independence should be $P(A|B) = P(A)$ or $P(B|A) = P(B)$. But to include that case of $P(A) = 0$ or $P(B) = 0$, the mathematical definition of independence is $P(A \cap B) = P(A)P(B)$, which is equivalent to $P(A^c \cap B) = P(A^c)P(B)$ or $P(A \cap B^c) = P(A)P(B^c)$ or $P(A^c \cap B^c) = P(A^c)P(B^c)$. The definition is extended in the following to independence between n events.

Definition Events $A_1, ..., A_n$ are called *independent* if $P(\cap_{i=1}^n B_i) = \prod_{i=1}^n P(B_i)$ where B_i is A_i or A_i^c. Events $A_1, ..., A_n$ are called *pairwise independent* if any pair of two events are independent.

The above definition implies, if $A_1, ..., A_n$ are independent (pairwise independent), then $A_{i_1}, ..., A_{i_k}$ are independent (pairwise independent). (Please DIY).

The σ-algebra generated by a single set A, denoted as $\sigma(A)$ is $\{\emptyset, A, A^c, \Omega\}$. Independence between $A_1, ..., A_n$ can be interpreted as independence between the σ-algebras: $\sigma(A_i), i = 1, ..., n$.

(ii). Borel-Cantelli Lemma.

The Borel-Contelli Lemma is considered as *sine qua non* of probability theory and is instrumental in proving the law of large numbers. Please note in the proof below the technique of using the indicator functions to handle probability of sets,

Theorem 4.1. (Borel-Contelli Lemma) For events $A_1, A_2, ...,$

$$(1) \quad \sum_{n=1}^{\infty} P(A_n) < \infty \Longrightarrow P(A_n, i.o.) = 0;$$

$$(2) \qquad \text{If } A_n \text{ are independent,} \quad \sum_{n=1}^{\infty} P(A_n) = \infty \Longrightarrow P(A_n, i.o.) = 1.$$

Here $A_n, i.o.$ means A_n happens infinitely often, i.e., $\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$.

Proof. (1): Let 1_{A_n} be the indicator function of A_n. Then, $A_n, i.o.$ is the same as $\sum_{n=1}^{\infty} 1_{A_n} = \infty$. Hence,

$$E(\sum_{i=1}^{\infty} 1_{A_n}) = \sum_{n=1}^{\infty} E1_{A_n} = \sum_{n=1}^{\infty} P(A_n) < \infty.$$

It implies $\sum_{i=1}^{n} 1_{A_n} < \infty$ with probability 1. This is equivalent to $P(A_n, i.o.) = 0$.

(2). $\sum_{n=1}^{\infty} P(A_n) = \infty$ implies $\prod_{k=n}^{\infty}(1 - P(A_k)) = 0$ since $\log(1 - x) \leq -x$ for $x \in [0, 1]$. for all $n \geq 1$. By dominated convergence theorem

$$E(\liminf 1_{A_n^c}) = E(\lim_n \prod_{k=n}^{\infty} 1_{A_k^c}) = \lim_n E(\prod_{k=n}^{\infty} 1_{A_k^c}) = \lim_n \prod_{k=n}^{\infty}(1 - P(A_k)) = 0.$$

Then, $P(\liminf_n A_n^c) = 0$ and hence $P(\limsup A_n) = 1$. □

As an immediate consequence,

Corollary (BOREL'S 0-1 LAW) If $A_1, ..., A_n, ...$ are independent, then $P(A_n, i.o.) = 1$ or 0 according as $\sum_n P(A_n) = \infty)$ or $< \infty$.

Even though the above 0-1 law appears to be simple, its impact and implication is profound. More generally, suppose $A \in \cap_{n=1}^{\infty}\sigma(A_j, j \geq n)$, the so-called *tail σ-algebra. A* is called a *tail event.* Then, the independence of $A_1, ..., A_n, ...$ implies $P(A) = 0$ or 1. The key fact here is that A is independent of A_n for any $n \geq 1$, such as, for example, $\{A_n, i.o.\}$ or $\{\sum_{i=1}^n 1_{A_i}/\log(n) \to \infty\}$. A more general result involving independent random variables to be introduced below is the Kolmogorov's 0-1 law to be introduced later.

The following example can be viewed as a strengthening of the Borel-Cantelli lemma.

EXAMPLE 4.1 Suppose $A_1, ..., A_n, ...$ are independent events with $\sum_n p_n = \infty$ where $p_n = P(A_n)$. Then,

$$X_n \equiv \frac{\sum_{i=1}^n 1_{A_i}}{\sum_{i=1}^n p_i} \to 1 \qquad a.s..$$

Proof Since

$$E(X_n - 1)^2 = \frac{\sum_{i=1}^n p_i(1-p_i)}{(\sum_{i=1}^n p_i)^2} \leq \frac{1}{\sum_{i=1}^n p_i} \to 0,$$

it follows that $X_n \to 1$ in L_2 and therefore also in probability by the Chebyshev inequality:

$$P(|X_n - 1| > \epsilon) \leq \frac{E(X_n-1)^2}{\epsilon^2} \leq \frac{1}{\epsilon^2 \sum_{i=1}^n p_i} \to 0.$$

Consider $n_k \uparrow \infty$ as $k \to \infty$, such that

$$\sum_{k=1}^{\infty} \frac{1}{\sum_{i=1}^{n_k} p_i} < \infty \qquad \text{and} \qquad \frac{\sum_{i=1}^{n_{k+1}} p_i}{\sum_{i=1}^{n_k} p_i} \to 1.$$

Then,

$$\sum_{i=1}^{\infty} P(|X_{n_k} - 1| > \epsilon) < \infty.$$

The Borel-Cantelli lemma implies $X_{n_k} \to 1$ a.s.. Observe that, for $n_k \leq n \leq n_{k+1}$,

$$1 \leftarrow \frac{\sum_{i=1}^{n_k} 1_{A_i}}{\sum_{i=1}^{n_{k+1}} p_i} \leq X_n = \frac{\sum_{i=1}^n 1_{A_i}}{\sum_{i=1}^n p_i} \leq \frac{\sum_{i=1}^{n_{k+1}} 1_{A_i}}{\sum_{i=1}^{n_k} p_i} \to 1, \quad a.s..$$

The desired convergence holds. □

Remark. The trick of bracketing X_n by the two quantities in the above inequality is also used in proving the uniform convergence of the empirical distribution to the population distribution:

$$|F_n(x) - F(x)| \to 0, \qquad a.s.,$$

where $F_n(x) = (1/n)\sum_{i=1}^n 1_{\{\xi_i \leq x\}}$ and ξ_i are iid with cdf F. The idea is further elaborated in the context of empirical approximation in terms of bracketing/packing numbers.

EXAMPLE 4.2. Repeatedly toss a coin, which has probability p to be head and $q = 1-p$ to be tail on each toss. Let $X_n = H$ or T when n-th toss is a head or tail. Let

$$l_n = \max\{m \geq 0 : X_n = H, X_{n+1} = H, ..., X_{n+m-1} = H, X_{n+m} = T\}$$

be the length of run of heads starting from n-th toss. Then,

$$\limsup_n l_n/\log n = 1/\log(1/p).$$

Proof. l_n follows a geometric distribution, i.e.,

$$P(l_n = k) = qp^k, \qquad P(l_n \geq k) = P(X_n = 1, ..., X_{n+k-1} = 1) = p^k \qquad k = 0, 1, 2, ...$$

For any $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P\Big(l_n > (1+\epsilon)\frac{\log n}{\log(1/p)}\Big) \le \sum_{n=1}^{\infty} p^{(1+\epsilon)\frac{\log n}{\log(1/p)}} \le \sum_{n=1}^{\infty} e^{-(1+\epsilon)\log n} = \sum_{n=1}^{\infty} n^{-(1+\epsilon)} < \infty$$

By the Borel-Cantelli lemma,

$$\limsup_n \frac{l_n}{\log n/\log(1/p)} \le 1.$$

We next try to find a subsequence with limit as large as 1. . Let d_n be the integer part of $\log n/\log(1/p)$ and let $r_n = \sum_{i=1}^{n} d_i$. Then $r_n \approx n\log n/\log(1/p)$ and $\log(r_n) \approx \log(n)$. Set

$$A_n = \{X_{r_n} = H, X_{r_n+1} = H, ..., X_{r_n+d_n-1} = H\}$$

Then $A_n, n \ge 1$ are independent, and

$$P(A_n) = p^{d_n} = e^{d_n \log p} \approx 1/n$$

Therefore, $\sum_n P(A_n) = \infty$. It then follows from the Borel Cantelli lemma that $P(A_n, i.o,) = 1$. Since $A_n = \{l_{r_n} \ge d_n\}$, we have

$$\limsup_n \frac{l_n}{\log n/\log(1/p)} \ge \limsup_n \frac{l_{r_n}}{\log(r_n)/\log(1/p)} = \limsup_n \frac{l_{r_n}}{d_n} \ge 1.$$

□

Remark. An analogous problem occurs in the setting of Poisson processes. Consider a Poisson process with intensity $\lambda > 0$. The sojourn times (time between two consecutive events) $\xi_0, \xi_1, ...$ are iid $\sim$ exponential distribution with mean $1/\lambda$. Then, $\limsup_{x\to\infty} l_x/x = 1/\lambda$, where l_x the time period between x and the time of the event right after x.

(iii). *Independence between σ-algebras and between random variables.*

Definitions. Let $\mathcal{A}_1, ..., \mathcal{A}_n$ be σ-algebras. They are called independent if $A_1, ..., A_n$ are independent for any $A_j \in \mathcal{A}_j$, $j = 1, ..., n$. Random variables $X_1, ..., X_n$ are called independent, if the σ-algebras generated by X_j, $1 \le j \le n$, are independent, i.e.,

$$P(\cap_{j=1}^{n} X_j^{-1}(B_j)) = \prod_{j=1}^{n} P(X_j^{-1}(B_j)) \qquad \text{or} \qquad P(X_1 \in B_1, ..., X_n \in B_n) = \prod_{j=1}^{n} P(X_j \in B_j)$$

for any Borel sets $B_1, ..., B_n$ in $(-\infty, \infty)$.

There are several equivalent definition of the independence of random variables:

Two r.v.s X and Y are called independent, if $E(g(X)f(Y)) = E(g(X))E(f(Y))$ for all bounded (measurable) functions g and f. or, equivalently, if

$$P(X \le t, \text{ and } Y \le s) = \prod_{i=1}^{n} P(X_j \le t_j) \qquad \text{for all } t_j \in (-\infty, \infty),\ j = 1, ..., n.$$

i.e., in terms of cumulative distribution functions.

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \qquad \text{for all } x, y.$$

If the joint density exists, This is the same as $f_{X,Y}(x, y) = f_X(x)f_X(y)$.

Roughly speaking, independence between two r.v.s X and Y is interpreted as X taking any value "has nothing to do with" Y taking any value, and vice versus.

(iv). *Conditional expectation.*

(1). Conditional distribution and conditional expectation with respect to a set A.

Suppose A is a set with $P(A) > 0$, and X is a random variable. Then, the *conditional expectation* is

$$E(X|A) \equiv E(X1_A)/P(A).$$

The *conditional distribution* of X given A is

$$P(X \leq t|A) = P(\{X \leq t\} \cap A)/P(A)$$

Then, $E(X|A) = \int t dP(X \leq t|A)$, if exist.

As a simple example, let $X \sim Unif[0,1]$. Let $A_i = \{i - 1/n < X \leq i/n\}$ for $i = 1, ..., n$.

$$E(X|A_i) \equiv E(X1_{A_i})/P(A_i) = (i - 1/2)/n.$$

Similarly $E(X|A_i^c) \equiv E(X1_{A_i^c})/P(A_i^c)$.

Interpretation: *$E(X|A)$ is the weighted "average" (expected value) of X over the set A.*

(2). Conditional expectation with respect to a r.v..

For two random variables X, Y, $E(X|Y)$ is a function of Y, i.e., measurable to $\sigma(Y)$, such that, for any $A \in \sigma(Y)$,

$$E(X1_A) = E[E(X|Y)1_A].$$

Interpretation: *$E(X|Y)$ is the weighted "average" (expected value) of X over the set $\{Y = y\}$ for all y. It is a function of Y and therefore is a r.v. measurable to $\sigma(Y)$.*

If their joint density $f(x, y)$ exists, then the conditional density of X given $Y = y$ is $f_{X|Y}(x|y) \equiv f(x, y)/f_Y(y)$. And

$$E(X|Y = y) \equiv \int x f_{X|Y}(x|y) dx.$$

(3). Conditional expectation with respect to a σ-algebra $\mathcal{A}$.

Conditional expectation w.r.t. a σ-algebra is the most fundamental concept in probability theory, especially in martingale theory in which the very definition of martingale depends on conditional expectation.

Recall that a random variable, say X, is measurable to a σ-algebra $\mathcal{A}$ is that for any interval (a, b), $\{\omega : X(\omega) \in (a, b)\} \in \mathcal{A}$. In other words, $\sigma(X) \subseteq \mathcal{A}$ is interpreted as all information about X, (which is $\sigma(X)$), is contained in $\mathcal{A}$.

If $\mathcal{A} = \sigma(A_1, ..., A_n)$ where $A_i \cap A_j = \emptyset$, then X measurable to $\mathcal{A}$ implies X must be constant over each A_i. If $\mathcal{A}$ is generated by a r.v. Y, then X measurable to $\mathcal{A}$ implies ξ must be a function of Y. A heuristic understanding is that if Y is known, then there is no uncertainty of X, or if Y assumes one value, X cannot assume more than one values.

Definition For a random variable X and a completed σ-algebra $\mathcal{A}$, $E(X|\mathcal{A})$ is defined as an $\mathcal{A}$-measurable random variable such that, for any $A \in \mathcal{A}$,

$$E(X1_A) = E(E(X|\mathcal{A})1_A),$$

i.e. $E(X|A) = E(E(X|\mathcal{A})|A)$ for every $A \in \mathcal{A}$ with $P(A) > 0$.

If $\mathcal{A} = \sigma(A_1, ..., A_n)$ where $A_i \cap A_j = \emptyset$, then

$$E(X|\mathcal{A}) = \sum_{j=1}^{n} E(X|A_i)1_{A_i},$$

which is a r.v. that, on each A_i, takes the conditional average of X, i.e., $E(X|A_i)$, as its value. Motivated from this simple case, we may obtain an important understanding of the conditional

expectation X w.r.t. a σ-algebra $\mathcal{A}$: *a new r.v. as the "average" of the r.v. X on each "unsplitable" or "smallest" set of the σ-algebra $\mathcal{A}$.*

Conditional mean/expectation with respect to σ algebra shares many properties just like the ordinary expectation.

Properties:

(1). $E(aX + bY|\mathcal{A}) = aE(X|\mathcal{A}) + bE(Y|\mathcal{A})$

(2). If $X \in \mathcal{A}$, then $E(X|\mathcal{A}) = X$.

(4). $E(E(X|\mathcal{F})|\mathcal{A}) = E(X|\mathcal{A})$ for two σ-algebras $\mathcal{A} \subseteq \mathcal{F}$.

Further properties, such as the dominated convergence theorem, Fatou's lemma and monotone convergence theorem also hold for conditional mean w.r.t. a σ-algebra. (See DIY exercises.)

(v). Kolmogorov's 0-1 law.

One of the most important theorem in probability theory is the *martingale convergence theorem*. In the following, we provide a simplified version, without a rigorous introduction of martingale and without giving a proof.

Theorem 1.2 (SIMPLIFIED VERSION OF MARTINGALE CONVERGENCE THEOREM) Suppose $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for $n \geq 1$. Let $\mathcal{F} = \sigma(\cup_{n=1}^{\infty} F_n)$. For any random variable X with $E(|X|) < \infty$,

$$E(X|\mathcal{F}_n) \to E(X|\mathcal{F}), \qquad a.s.$$

The martingale convergence theorem, even with the simplified version, has broad applications. For example, One of the most basic 0-1 laws: the Kolomogorov 0-1 law, can be established upon it.

Corollary (KOLOMOGOROV 0-1 LAW) Suppose $X_1, ..., X_n, ...$ are a sequence of independent r.v.s. Then all tails events are have probability 0 or 1.

Proof. Suppose A is a tail event. Then A is independent of $X_1, ..., X_n$ for any fixed n. Therefore $E(1_A|\mathcal{F}_n) = P(A)$ where $\mathcal{F}_n$ is the σ-algebra generated by $X_1, ..., X_n$. But, by Theorem 1.2, $E(1_A|\mathcal{F}_n) \to 1_A$ a.s.. Hence $1_A = P(A)$, and A can only be 0 or 1. □

A heuristic interpretation of Kolmogorov's 0-1 law could be in the perspective of information. When σ-algebras $\mathcal{A}_1, ..., \mathcal{A}_n, ...$ are independent, the information carried by each $\mathcal{A}_i$ are independent or unrelated or non-overlapping. Then, the information carried by $\{\mathcal{A}_n, \mathcal{A}_{n+1}, ...\}$ shall shrink to 0 as $n \to \infty$, as, if otherwise, $\mathcal{A}_n, \mathcal{A}_{n+1}, ...$ would have something in common.

As straightforward applications of Kolmogorov's 0-1 law:

Corollary Suppose $X_1, ..., X_n, ...$ are a sequence of independent random variables. Then,

$$\liminf_n X_n, \quad \limsup_n X_n, \quad \limsup_n S_n/a_n \quad \text{and} \quad \liminf_n S_n/a_n$$

must be either a constant or ∞ or $-\infty$, a.s., where $S_n = \sum_{i=1}^n X_i$ and $a_n \uparrow \infty$.

Proof. Consider $A = \{\omega : \liminf_n X_n(\omega) > a\}$. Try to show A is a tail event. (DIY). □

Remark. Without invoking martingale convergence theorem, Kolmogorov's 0-1 law can be shown through $\pi - \lambda$ theorem, which we do not plan to cover.

DIY EXERCISES.

Exercise 4.1 ⋆⋆ Suppose X_n are iid random variables. Then $X_n/n^{1/p} \to 0$ a.s. if and only if $E(|X_n|^p) < \infty$ for $p > 0$. Hint: Borel-Cantelli lemma.

Exercise 4.2 ⋆⋆⋆ Let X_n be iid r.v.s with $E(X_n) = \infty$. Show that $\limsup_n |S_n|/n = \infty$ a.s. where $S_n = X_1 + \cdots + X_n$.

Exercise 4.3 ⋆⋆⋆ Suppose X_n are iid nonnegative random variables such that $\sum_{k=1}^{\infty} kP(X_1 > a_k) < \infty$ for $a_k \uparrow \infty$. Show that $\limsup_n \max_{1 \leq i \leq n} X_i/a_n \leq 1$ a.s.

Exercise 4.4 ⋆⋆⋆⋆ (Empirical Approximation) For every fixed $t \in [0,1]$, $S_n(t)$ is a sequence of random variables such that, with probability 1 for some $p > 0$,

$$|S_n(t) - S_n(s)| \leq n|t-s|^p,$$

for all $n \geq 1$ and all $t, s \in [0,1]$. Suppose for every constant $C > 0$, there exists an $c > 0$ such that

$$P(|S_n(t)| > C(n \log n)^{1/2}) \leq e^{-cn} \quad \text{for all } n \geq 1 \text{ and } t \in [0,1].$$

Show that, for any $p > 0$,

$$\frac{\max\{|S_n(t)| : t \in [0,1]\}}{(n \log n)^{1/2}} \to 0 \qquad a.s..$$

Hint: Borel-Cantelli lemma.

Chapter 5. Weak law of large numbers.

For a sequence of independent r.v.s $X_1, X_2, ...$, classical law of large numbers is typically about the convergence of partial sums

$$\frac{S_n - E(S_n)}{n} = \frac{\sum_{i=1}^n [X_i - E(X_i)]}{n},$$

where $S_n = \sum_{i=1}^n X_i$ here and throughout this Chapter. A more general form is the convergence of

$$\frac{S_n - a_n}{b_n}$$

for some constants a_n and b_n. Weak law is convergence in probability and strong law is convergence a.s..

The following proposition may be called L^2 weak law of large numbers which implies the weak law of large numbers.

PROPOSITION *Suppose $X_1, ..., X_n, ...$ are iid with mean μ and finite variance σ^2. Then,*

$$S_n/n \to \mu \qquad \textit{in probability and in } L^2.$$

Proof. Write

$$E(S_n/n - \mu)^2 = (1/n)\sigma^2 \to 0.$$

Therefore L^2 convergence holds. And convergence in probability is implied by the Chebyshev inequality. □

The above proposition implies that classical weak law of large numbers holds quite trivially in a standard setup with the r.v.s being iid with finite variance. In fact, in such a standard setup strong law of large numbers also holds, as to be shown in Chapter 6. However, the fact that convergence in probability is implied in L^2 convergence plays a central role is establishing weak law of large numbers. For a example, a straightforward extension of the above proposition can be:

For independent r.v.s $X_1, ...,$, $(S_n - E(S_n))/b_n \to 0$ in probability if $(1/b_n^2)\sum_{i=1}^n \text{var}(X_i) \to 0$, for some $b_n \uparrow \infty$.

The following theorem about general weak law of large numbers is a combination of the above extension and the technique of truncation.

Theorem 5.1. WEAK LAW OF LARGE NUMBERS *Suppose $X_1, X_2, ...$ are independent. Assume*
(1). $\sum_{i=1}^n P(|X_i| > b_n) \to 0$,
(2). $b_n^{-2} \sum_{i=1}^n E(X_i^2 1_{\{|X_i| \le b_n\}}) \to 0$,
where $0 < b_n \uparrow \infty$. Then $(S_n - a_n)/b_n \to 0$ in probability, where $a_n = \sum_{j=1}^n E(X_i 1_{\{|X_i| \le b_n\}})$.

Proof. Let $Y_j = X_j 1_{\{|X_j| \le b_n\}}$. Consider

$$\frac{\sum_{j=1}^n Y_j - a_n}{b_n} = \frac{\sum_{j=1}^n [Y_j - E(Y_j)]}{b_n},$$

which is mean 0 and converges to 0 in L^2 by (2). Therefore it also converges to 0 in probability. Notice that

$$P\Big(\frac{S_n - a_n}{b_n} = \frac{\sum_{j=1}^n Y_j - a_n}{b_n}\Big) = P(S_n = \sum_{j=1}^n Y_j)$$

$$
\begin{aligned}
&\geq P(X_j = Y_j \text{ for all } 1 \leq j \leq n) = \prod_{j=1}^{n} P(X_j = Y_j) \qquad \text{by independence} \\
&= \prod_{j=1}^{n} P(|X_j| \leq b_n) = \prod_{j=1}^{n} [1 - P(|X_j| > b_n)] = e^{\sum_{j=1}^{n} \log[1 - P(|X_j| > b_n)]} \\
&\approx e^{-\sum_{j=1}^{n} P(|X_j| > b_n)} \\
&\to 1 \qquad \text{by (1).}
\end{aligned}
$$

Hence $(S_n - a_n)/b_n \to 0$ in probability. □

Theorem 5.2. *Suppose $X, X_1, X_2, ...$ are iid. Then, $S_n/n - \mu_n \to 0$ in probability for some μ_n, if and only if*

$$xP(|X_1| > x) \to 0 \qquad \text{as } x \to \infty.$$

in which case $\mu_n = E(X1_{\{|X| \leq n\}}) + o(1)$.

Proof. "$\Longleftarrow$" Let $a_n = n\mu_n$ and $b_n = n$ in Theorem 5.1. Condition (1) follows. To check Condition (2), write, as $n \to \infty$,

$$
\begin{aligned}
& b_n^{-2} \sum_{i=1}^{n} E(X_i^2 1_{\{|X_i| \leq b_n\}}) = \frac{1}{n} E(X^2 1_{\{|X| \leq n\}}) \leq \frac{1}{n} E(\min(|X|, n)^2) \\
= \; & \frac{1}{n} \int_0^{\infty} 2xP(\min(|X|, n) > x)dx = \frac{1}{n} \int_0^{n} 2xP(|X| > x)dx \\
= \; & \frac{1}{n} \int_M^{n} 2xP(|X| > x)dx + o(1) \qquad \text{for any fixed } M > 0 \\
= \; & \frac{2}{n} \int_M^{n} xP(|X| > x)dx + o(1) \leq 2 \sup_{x \geq M} xP(|X| > x) + o(1),
\end{aligned}
$$

as $n \to \infty$. Since M is arbitray, Condition (2) holds. And the WLLN follows from Theorem 1.3.

"$\Longrightarrow$" Let $X^*, X_1^*, ...$ be iid following the same distribution of X and are independent of $X, X_1, ...$ Set $\xi_i = X_i - X_i^*$ (symmetrization) and $\tilde{S}_n = \sum_{i=1}^{n} \xi_i$. Then, $\tilde{S}_n/n \to 0$ in probability. The Levy inequality in Exercise 5.1 implies $\max\{|\tilde{S}_j| : 1 \leq j \leq n\}/n \to 0$ in probability, which further ensures $\max\{|\xi_j| : 1 \leq j \leq n\}/n \to 0$ in probability. For any $\epsilon > 0$,

$$
\begin{aligned}
& nP(|X| \geq n\epsilon)P(|X^*| \leq .5n\epsilon) = nP(|X| \geq n\epsilon, |X^*| \leq .5n\epsilon) \leq nP(|X - X^*| \geq .5n\epsilon) \\
\approx \; & 1 - [1 - P(|X - X^*| \geq .5n\epsilon)]^n = P(\max_{1 \leq j \leq n} |\xi_j| > .5n\epsilon) \to 0.
\end{aligned}
$$

As a result, for any $\epsilon > 0$,

$$nP(|X| \geq n\epsilon) \approx nP(|X| \geq n\epsilon)[1 - P(|X| \geq .5n\epsilon)] \to 0,$$

which is equivalent to $xP(|X| > x) \to 0$ as $x \to \infty$. □

EXAMPLE 5.1. Suppose $X_1, X_2, ...$ are i.i.d. with common density f symmetric about 0 and c.d.f such that $1 - F(t) = 1/(t \log t)$, for $t > 3$. Then, $S_n/n \to 0$ in probability. But $S_n/n \nrightarrow 0$, a.s..

The convergence in probability is a consequence of Theorem 5.2 with $\mu_n = 0$ and checking the condition $xP(|X| > x) \to 0$ as $x \to \infty$. The convergence a.s. is untrue because $X_n/n \nrightarrow 0$ a.s. by Borel-Cantelli lemma. □

Corollary. *Suppose $X_1, ..., X_n, ...$ are i.i.d. with $E(|X_i|) < \infty$. Then, $S_n/n \to E(X_1)$ in probability.*

Proof. Since, as $x \to \infty$,

$$xP(|X_i| > x) = o(1)\int_0^x P(|X_i| > t)dt = o(1)\int_0^\infty P(|X_i| > t)dt = o(1)E(|X_i|),$$

the WLLN follows from Theorem 5.2. □

Example 5.2. The St. Petersberg Paradox. Let $X, X_1, ..., X_n, ...$ be iid with $P(X = 2^k) = 2^{-k}$, $k = 1, 2, ...$ Then, $E(X) = \infty$ and

$$\frac{S_n}{n \log n} \to \frac{1}{\log 2} \qquad \text{in probability.}$$

Proof. Notice that $P(X \geq 2^k) = 2^{-k+1}$. Let $k_n \approx \log\log n/\log 2$, $m_n = \log n/\log 2 + k_n$ and $b_n = 2^{m_n} = 2^{k_n} n \approx n \log n$. m_n is an integer. Then,

$$nP(X \geq b_n) = n2^{-m_n+1} \approx 2n/n \cdot 2^{-k_n} \to 0.$$

And

$$E(X^2 1_{\{|X| \leq b_n\}}) = \sum_{k=1}^{m_n} 2^{2k} 2^{-k} = \sum_{k=1}^{m_n} 2^k \leq 2 \times 2^{m_n} = 2b_n.$$

Then,

$$\frac{nE(X^2 1_{\{|X| \leq b_n\}})}{b_n^2} \leq \frac{2nb_n}{b_n^2} = \frac{2n}{b_n} = \frac{2n}{2^{m_n}} = \frac{2n}{n2^{k_n}} \to 0.$$

Let $a_n = nE(X1_{\{|X| \leq b_n\}})$.

$$a_n = n\sum_{k=1}^{m_n} 2^k 2^{-k} = nm_n = n \log n/\log 2 + nk_n \approx b_n \log 2.$$

The desired convergence is implied by Theorem 5.2. □

Example 5.3. "Unfair fair game". You pay one dollar to buy a lottery. The lottery has infinite number of numbered balls. If number k occurs, you are paid by 2^k dollars. The number k ball occurs with probability

$$p_k \equiv \frac{1}{2^k k(k+1)}.$$

Is this a fair game?

In a sense, it is fair. Let X be gain/loss of the outcome. Then $P(X = 2^k - 1) = p_k, k = 1, 2, ...$ and $P(X = -1) = 1 - \sum_k p_k$. Then $E(X) = 0$.

If one buys the lottery on daily basis, one time every day. Let X_n be gain/loss of day n and S_n be the cumulative gain/loss up to day n. Then,

$$\frac{S_n}{n/\log n} \to -\log 2 \qquad \text{in probability,}$$

meaning that in the long time, he/she is nearly certainly in red. □

Example 5.4. Compute the limit of

$$\int_0^1 \cdots \int_0^1 \frac{x_1^2 + \cdots x_1^2 +}{x_1 + \cdots + x_n} dx_1 \cdots dx_n.$$

Solution. The above integral is the same as

$$E\Big(\frac{X_1^2 + \cdots + X_n^2}{X_1 + \cdots + X_n}\Big),$$

where $X_1, ..., X_n, ...$ are iid $\sim Unif[0,1]$. Since, by the WLLN

$$(1/n)\sum_{i=1} X_i^2 \to E(X_1^2) = \int_0^1 x^2 dx = 1/3 \quad \text{and} \quad (1/n)\sum_{i=1} X_i \to E(X_1) = 1/2,$$

with the convergence being convergence in probability, we have

$$\frac{X_1^2 + \cdots + X_n^2}{X_1 + \cdots + X_n} \to 2/3 \qquad \text{in probability.}$$

The r.v. on the left hand side is bounded by 1. By the dominated convergence, its mean also converges to 2/3. Then the limit of the integral is 2/3. □

REMARK. The following WLLN for array of r.v.s. is a slight generalization of Theorem 5.1.

Suppose $X_{n,1}, ..., X_{n,n}$ are independent r.v.s. If

$$\sum_{i=1}^n P(|X_{n,i}| > b_n) \to 0 \quad \text{and} \quad (1/b_n^2)\sum_{i=1}^n E(X_{n,i}^2 1_{\{|X_{n,i}| \le b_n\}}) \to 0,$$

Then,

$$\frac{\sum_{i=1}^n X_{n,i} - a_n}{b_n} \to 0 \qquad \text{in probability}$$

where $a_n = \sum_{i=1}^n E(X_{n,i} 1_{\{|X_{n,i}| \le b_n\}})$.

DIY EXERCISES.

Exercise 5.1 (LEVY'S INEQUALITY) Suppose $X_1, X_2, ...$ are independent and symmetric about 0. Then,

$$P(\max_{1 \le j \le n} |S_j| \ge \epsilon) \le 2P(|S_n| \ge \epsilon)$$

Exercise 5.2 Show $S_n/(n \log n) \to -\log 2$ in probability in Example 5.4. Hint: Choose $b_n = 2^{m_n}$ with $m_n = \{k : 2^{-k} k^{-3/2} \le 1/n\}$ and proceed as in Example 5.2.

Exercise 5.3 For Example 1.4, prove that $S_n/b_n \to 0$ in probability, if $b_n/(n/\log n) \uparrow \infty$.

Exercise 5.4 (MARCINKIEWICZ-ZYGMUND WEAK LAW OF LARGE NUMBERS) Suppose $x^p P(|X| > x) \to 0$ as $x \to \infty$ for some $0 < p < 2$. Prove that

$$\frac{S_n - nE(X 1_{\{|X| \le n^{1/p}\}})}{n^{1/p}} \to 0 \qquad \text{in probability.}$$

Chapter 6. Strong law of large numbers.

For r.v.s $X_1, X_2, ...$, convergence of series means the convergence of its partial sums $S_n = \sum_{i=1}^n X_i$, as $n \to \infty$. We shall denote the convergence of S_n a.s. just as $\sum_{n=1}^\infty X_n < \infty$ a.s.. The following Kolmogorov inequality is the key to establishing a.s. convergence of series for independent r.v.s.

(i). Kolmogorov inequality.

Theorem 6.1. KOLMOGOROV INEQUALITY *Suppose $X_1, X_2, ..., X_n$ are independent with $E(X_i) = 0$ and $\text{var}(X_i) < \infty$. $S_j = X_1 + ... + X_j$. Then,*

$$P(\max_{1\le j\le n} |S_j| \ge \epsilon) = \frac{\text{var}(S_n)}{\epsilon^2}.$$

Proof. Let $T = \min\{j \le n : |S_j| \ge \epsilon\}$, with minimum of empty set being ∞, i.e., $T = \infty$ $|S_j| \le \epsilon$ for all $1 \le j \le n$. Then, $\{T \le j\}$ *or* $\{T = j\}$ *only depends on* $X_1, ..., X_j$. And, as a result,

$$\{T \ge j\} = \{T \le j-1\}^c = \{S_i \le \epsilon, 1 \le i \le j-1\}$$

only depends on $X_1, ..., X_{j-1}$ and therefore is independent of $X_j, X_{j+1}, ...$ Write

$$\begin{aligned}
& P(\max_{1\le j\le n} |S_j| \ge \epsilon) = P(T \le n) \ \le \ \epsilon^{-2} E(|S_T|^2 1_{\{T\le n\}}) \ \le \ \epsilon^{-2} E(|S_{T\wedge n}|^2) \\
= \ & \epsilon^{-2} E(|\sum_{j=1}^{T\wedge n} X_j|^2) = \epsilon^{-2} E(|\sum_{j=1}^{n} X_j 1_{\{T\ge j\}}|^2) \\
= \ & \epsilon^{-2}\Big\{ E(\sum_{j=1}^n X_j^2 1_{\{T\ge j\}}) + 2 \sum_{1\le i<j\le n}^n E(X_j X_i 1_{\{T\ge j\}} 1_{\{T\ge i\}} \Big\} \\
= \ & \epsilon^{-2}\Big\{ \sum_{j=1}^n E(X_j^2) P(T \ge j) + 2 \sum_{1\le i<j\le n}^n E(X_j) E(X_i 1_{\{T\ge j\}} 1_{\{T\ge i\}}) \Big\} \\
= \ & \epsilon^{-2} \sum_{j=1}^n E(X_j^2) P(T \ge j) + 0 \\
\le \ & \text{var}(S_n)/\epsilon^2.
\end{aligned}$$

□

EXAMPLE 6.1. (Extension to continuous time process.) Suppose $\{S_t : t \in [0, \infty)\}$ is a process with increments that are independent, zero mean and finite variance. If the path of S_t is right continuous, e.g.

$$P\Big(\max_{t\in[0,\tau]} |S_t| > \epsilon\Big) \le \frac{\text{var}(S_\tau)}{\epsilon^2}.$$

The examples of such processes are, e.g., compensated Poisson process and Brownian Motion.

Kolmogorov's inequality will later on be seen as a special case of martingale inequality. In the proof of Kolmogorov inequality, we have used a *stopping time* T, which is a r.v. associated with a process S_n or, more generally, a filtration, such that $T = k$ only depends on past and current values of the process: $S_1, ..., S_k$. Stopping time is one of the most important concepts and tools in martingale theory or stochastic processes.

(ii). Khintchine-Kolmogorov convergence theorem.

Theorem 6.2. (KHINTCHINE-KOLMOGOROV CONVERGENCE THEOREM) *Suppose $X_1, X_2, ...$ are independent with mean 0 such that $\sum_n \text{var}(X_n) < \infty$. Then, $\sum_n X_n < \infty$ a.s., i.e., S_n converges a.s. as well as in L^2 to $\sum_{n=1}^\infty X_n$.*

Proof. Define $A_{m,\epsilon} = \{\max_{j>m} |S_j - S_m| \le \epsilon\}$. Then, $\{\sum_{n=1}^{\infty} X_n < \infty\} = \cap_{\epsilon>0} \cup_m A_{m,\epsilon}$. By Kolmogorov's inequality

$$P(\max_{m<j\le n} |S_j - S_m| > \epsilon) \le \frac{\text{var}(S_n - S_m)}{\epsilon^2} = \frac{1}{\epsilon^2} \sum_{i=m+1}^{n} \text{var}(X_i) \le \frac{1}{\epsilon^2} \sum_{i=m+1}^{\infty} \text{var}(X_i).$$

By letting $n \to \infty$ first and then $m \to \infty$, we have

$$\lim_{m\to\infty} P(\max_{j>m} |S_j - S_m| > \epsilon) \to 0.$$

Then $\lim_m P(A_{m,\epsilon}) \to 1$. So $P(\cup_{m\ge 1} A_{m,\epsilon}) = 1$ for every $\epsilon > 0$. Hence,

$$P(\sum_n X_n < \infty) = P(\cap_{\epsilon>0} \cup_m A_{m,\epsilon}) = 1.$$

And a.s. convergence of S_n holds. Denote the a.s. limit as S_∞.

To show convergence of S_n in L^2, write

$$\begin{aligned} & E[(S_n - S_\infty)^2] = E[(S_n - \lim_k S_k)^2] = E[\lim_k (S_n - S_k)^2] \\ \le\ & \liminf_k E[(S_n - S_k)^2] \qquad \text{by Fatou's lemma} \\ =\ & = \liminf_k \sum_{j=n}^{k} \text{var}(X_j) = \sum_{j=n}^{\infty} \text{var}(X_j) \end{aligned}$$

which tends to 0, as $n \to \infty$. Therefore convergence in L^2 holds. □

EXAMPLE 6.2. Suppose $X_1, ...$ are iid with zero mean and finite variance. Then $\sum_n a_n X_n < \infty$ a.s. if and only if $\sum_n a_n^2 < \infty$.

"$\Longleftarrow$" is a direct consequence of Theorem 6.2.. "$\Longrightarrow$" follows from the central limit theorem to be shown in Chapter 8.

(iii). Kolmogorov three series theorem

For independent random variables, Kolmogorov three series theorem is the ultimate result in providing sufficient and necessary conditions for the convergence of series a.s..

Theorem 6.3. (KOLMOGOROV THREE SERIES THEOREM) *Suppose $X_1, X_2, ...$ are independent. Let $Y_n = X_n 1_{\{|X_n|\le 1\}}$ Then, $\sum_n X_n < \infty$ a.s. if and only if (1). $\sum_n P(|X_n| > 1) < \infty$; (2). $\sum_n E(Y_n) < \infty$; and (3). $\sum_n \text{var}(Y_n) < \infty$.*

Proof. "$\Longleftarrow$": The convergence of $\sum_n (Y_n - E(Y_n))$ is implied by (3) and Theorem 6.2. Together with (2), it ensures $\sum_n Y_n < \infty$ a.s.. On the other hand, Condition (1) and Borel-Cantelli lemma implies $P(X_n \ne Y_n, i.o.) = 0$. Consequently, $\sum_n X_n$ converges.

"$\Longrightarrow$" (An unconventional proof). It's straightforward that Condition (1) holds. Then $\sum_n Y_n < \infty$ a.s. since $P(X_n \ne Y_n,\ i.o.) = 1$. If condition (3) does not hold, by the central limit theorem to be shown in the next chapter,

$$\frac{1}{\sqrt{\sum_{i=1}^n \text{var}(Y_i)}} \sum_{i=1}^{n} [Y_i - E(Y_i)] \to N(0,1),$$

in distribution. Hence $P(|\sum_{i=1}^n Y_i| > M) \to 0$ as $n \to \infty$ for any fixed $M > 0$, which contradicts with $\sum_n Y_n < \infty$ a.s.. Hence condition (3) holds. Theorem 6.2 then ensures $\sum_n (Y_n - E(Y_n)) < \infty$ a.s.. As a result, $\sum_n E(Y_n) < \infty$ and condition (2) also holds. □

Remark. Suppose X_n is truncated at any constant $\epsilon > 0$ rather than 1 in Theorem 6.3, the theorem still holds.

Corollary. *Suppose $X, X_1, X_2, ...$ are iid with $E(|X|^p) < \infty$ for some $0 < p < 2$. Then, $\sum_{n=1}^{\infty}[X_n - E(X)]/n^{1/p} < \infty$ a.s. for $1 < p < 2$; and $\sum_{n=1}^{\infty} X_n/n^{1/p} < \infty$ a.s. for $0 < p < 1$.*

We leave the proof as Exercise 6.2.

Strong law of large numbers (SLLN) is a central result in classical probability theory. The convergence of series estabalished in Section 1.6 paves a way towards proving SLLN using the Kronecker lemma.

(iv). *Kronecker lemma and Kolmogorov's criterion of SLLN.*

Kronecker Lemma. Suppose $a_n > 0$ and $a_n \uparrow \infty$. Then $\sum_n x_n/a_n < \infty$ implies $\sum_{j=1}^n x_j/a_n \to 0$.

Proof. Set $b_n = \sum_{i=1}^n x_i/a_i$ and $a_0 = b_0 = 0$. Then, $b_n \to b_\infty < \infty$ and $x_n = a_n(b_n - b_{n-1})$. Write

$$\begin{aligned}
& \frac{1}{a_n}\sum_{j=1}^n x_j = \frac{1}{a_n}\sum_{j=1}^n a_j(b_j - b_{j-1}) = \frac{1}{a_n}\Big[\sum_{j=1}^n a_j b_j - \sum_{j=1}^n a_j b_{j-1}\Big] \\
= \quad & b_n + \frac{1}{a_n}\Big[\sum_{j=1}^{n-1} a_j b_j - \sum_{j=1}^n a_j b_{j-1}\Big] = b_n + \frac{1}{a_n}\Big[\sum_{j=1}^n a_{j-1} b_{j-1} - \sum_{j=1}^n a_j b_{j-1}\Big] \\
= \quad & b_n - \frac{1}{a_n}\sum_{j=1}^n b_{j-1}(a_j - a_{j-1}) \\
\to \quad & b_\infty - b_\infty = 0.
\end{aligned}$$

□

The following proposition is an immediate application of the Kronecker lemma and the Khintchine-Kolmogorov convergence of series.

Proposition (Kolmogorov's criterion of SLLN). Suppose $X_1, X_2...$, are independent such that $E(X_n) = 0$ and $\sum_n \text{var}(X_n)/n^2 < \infty$. Then, $S_n/n \to 0$ a.e..

Proof. Consider the series $\sum_{i=1}^n X_i/i < \infty$, $n \geq 1$. Then Theorem 6.2 implies $\sum_n X_n/n < \infty$ a.s.. And the above Kronecker Lemma ensures $S_n/n \to 0$ a.s.. □

Obviously, if $X, X_1, X_2, ...$ are iid with finite variance, the above proposition implies the SLLN: $S_n/n \to E(X)$ a.s.. In fact, a stronger result than the above SLLN is also straightforward:

Corollary. If $X_1, X_2, ...$ are iid with mean μ and finite variance. Then,

$$\frac{S_n - n\mu}{\sqrt{n(\log n)^\delta}} \to 0 \qquad .a.s.$$

for any $\delta > 1$.

We leave the proof as an exercise.

The corollary gives a rate of a.s. convergence of sample mean S_n/n to population mean μ at a rate $n^{-1/2}(\log n)^\delta$ with $\delta > 1/2$. This is, although not the sharpest rate, close to the sharpest rate of a.s. convergence at $n^{-1/2}(\log\log n)^{1/2}$ given in *Kolmogorov's law of iterated logarithm*:

$$\begin{cases} \limsup \frac{S_n - n\mu}{\sqrt{2\sigma^2 n \log\log n}} = 1 & a.s.. \\ \liminf \frac{S_n - n\mu}{\sqrt{2\sigma^2 n \log\log n}} = -1 & a.s.. \end{cases}$$

for iid r.v.s with mean μ and finite variance σ^2. We do not intend to cover the proofs of Kolmogorov's law of iterated logarithm.

(v) *Kolmogorov's strong law of large numbers.*

The above SLLN requires finite moments of the series. The most standard classical SLLN, established by Kolmogorov, for iid r.v.s. holds as long as the population mean exist. In statistical view, the sample mean shall always converge to the population mean as long as the population mean exists, without any further moment condition. In fact, the sample mean converges to a finite limit if and only if the population mean is finite, in which case, the limit is the population mean.

Theorem 6.4. *Kolmogorov's strong law of large numbers.* Suppose $X, X_1, X_2, ...$ are iid and $E(X)$ exists. Then,

$$S_n/n \to E(X), \quad a.s..$$

Conversely, if $S_n/n \to \mu$ which is finite, then $\mu = E(X)$.

Proof. Suppose first $E(X_1) = 0$. We shall utilize the above proposition of Kolmogorov's criterion of SLLN. Consider

$$Y_n = X_n 1_{\{|X_n| \le n\}} - E(X_n 1_{\{|X_n| \le n\}}).$$

Write

$$\begin{aligned} & \sum_{n=1}^{\infty} \frac{\text{var}(Y_n)}{n^2} \le \sum_{n=1}^{\infty} \frac{1}{n^2} E(X^2 1_{\{|X| \le n\}}) = E\Big(X^2 \sum_{n=1}^{\infty} \frac{1}{n^2} 1_{\{|X| \le n\}}\Big) \\ \le \quad & E\Big(X^2 \sum_{n \ge |X| \vee 1}^{\infty} \frac{2}{n(n+1)}\Big) \le 2E(|X|+1) < \infty \end{aligned}$$

It then follows from Kolmogorov's criterion of SLLN that

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \to 0 \qquad a.s..$$

Next, since $E(X_n 1_{\{|X_n| \le n\}}) \to E(X) = 0$. $\sum_{i=1}^{n} E(X_i 1_{\{|X_i| \le i\}})/n \to 0$. Hence,

$$\frac{1}{n} \sum_{i=1}^{n} X_i 1_{\{|X_i| \le i\}} \to 0 \qquad a.s..$$

Observe that $E(X) = 0$ implies $E|X| < \infty$, and

$$E|X| < \infty \Longleftrightarrow \sum_n P(|X| > n) < \infty \Longleftrightarrow P(|X_n| > n, i.o.) = 0 \Longleftrightarrow X_n/n \to 0 \quad a.s..$$

Therefore, $\sum_{i=1}^{n} X_i 1_{\{|X_i| > i\}}/n \to 0$ a.s.. As a result, the SLLN holds.

Suppose $E(X) < \infty$. the SLLN holds by considering $X_i - E(X)$, which is mean 0.

Suppose $E(X) = \infty$. Then, $(1/n)\sum_{i=1}^{n} X_i \wedge C \to E(X_1 \wedge C)$ a.s., which $\uparrow \infty$ when $C \uparrow \infty$. Since $S_n \ge \sum_{i=1}^{n} X_i \wedge C$, the SLLN holds. Likewise for the case $E(X) = -\infty$.

Conversely, if $S_n/n \to \mu$ a.s. where μ is finite, $X_n/n \to 0$ a.s.. Hence, $E|X| < \infty$ and $\mu = E(X)$ by the SLLN just proved. □

Remark Kolmogorov's SLLN also holds for r.v.s that are *pairwise independent* following the same distribution, which is slightly more general. We have chosen to follow the historic development of the classical probability theory.

(vi). *Strong law of large numbers when $E(X)$ does not exist.*

Kolmogorov's SLLN in Theorem 6.4 already shows that the classical SLLN does not hold if $E(X)$ does not exist, i.e., $E(X^+) = E(X^-) = \infty$. The SLLN becomes quite complicated. We introduce the theorem proved by W. Feller:

PROPOSITION Suppose $X, X_1, ...$ are iid with $E|X| = \infty$. Suppose $a_n > 0$ and a_n/n is nondecreasing. Then,

$$\begin{cases} \limsup |S_n|/a_n = 0 & \text{if } \sum_n P(|X| \geq a_n) < \infty \\ \limsup |S_n|/a_n = \infty & \text{if } \sum_n P(|X| \geq a_n) = \infty. \end{cases}$$

The proof is somewhat technical but still along the same line as the that of Kolmogorov's SLLN. Interested students may refer to the textbook by Durrett (page 67). We omit the details.

EXAMPLE 6.2. (THE ST. PETERSBURG PARADOX) See Example 1.5 in which we have shown

$$\frac{S_n}{n \log n} \to \frac{1}{\log 2} \qquad \text{in probability}$$

Analogous to the calculation therein,

$$\sum_{n=2}^{\infty} P(X \geq n \log n) = \sum_{n=2}^{\infty} P(X \geq 2^{\log(n \log n)/\log 2}) \geq \sum_{n=2}^{\infty} 2^{-\log(n \log n)/\log 2} = \sum_{n=2}^{\infty} 1/(n \log n) = \infty$$

By the above proposition,

$$\limsup \frac{S_n}{n \log n} = \infty \qquad a.s..$$

On the other hand, one can also show with same calculation that, for $\delta > 1$,

$$\limsup \frac{S_n}{n(\log n)^{\delta}} = 0 \qquad a.s..$$

□

The following Marcinkiewicz-Zygmund SLLN is useful in connecting the rate of convergence with the moments of the iid r.v.s.

Theorem 6.5. (MARCINKIEWICZ-ZYGMUND STRONG LAW OF LARGE NUMBERS). *Suppose $X, X_1, X_2, ...$ are iid and $E(|X|^p) < \infty$ for some $0 < p < 2$. Then,*

$$\begin{cases} \frac{S_n - nE(X)}{n^{1/p}} \to 0, \quad a.s. & \text{for } 1 \leq p < 2 \\ \frac{S_n}{n^{1/p}} \to 0 \quad a.s. & \text{for } 0 < p < 1. \end{cases}$$

Proof. The case with $p = 1$ is Kolmogorov's SLLN. The cases with $0 < p < 1$ and $1 < p < 2$ are consequences of the corollary following Theorem 1.6 and the Kronecker lemma. □

EXAMPLE 6.3 Suppose $X, X_1, X_2, ...$ are iid and X is symmetric with $P(X > t) = t^{-\alpha}$ for some $\alpha > 0$ and all large t.

(1). $\alpha > 2$: Then, $E(X^2) < \infty$, $S_n/n \to 0$ a.s. and, moreover, Kolmogorov's law of iterated logarithm gives the sharp rate of the a.s. convergence.

(2). $1 < \alpha \leq 2$: for any $0 < p < \alpha$

$$\frac{S_n}{n^{1/p}} \to 0, \quad a.s.$$

It implies that S_n/n converges to 0 a.s. at a rate faster than $n^{-1+1/p}$, but not at the rate of $n^{-1+1/\alpha}$. In particular, if $\alpha = 2$, S_n/n converges to $E(X)$ a.s. at a rate faster than $n^{-\beta}$ with any $0 < \beta < 1/2$, but not at the rate of $n^{-1/2}$.

(3). $0 < \alpha \leq 1$: $E(X)$ does not exist. For any $0 < p < \alpha$,

$$\frac{S_n}{n^{1/p}} \to 0, \quad a.s.$$

Moreover, the above proposition implies

$$\limsup \frac{|S_n|}{n^{1/\alpha}} = \infty \quad a.s. \quad \text{and} \quad \frac{S_n}{n^{1/\alpha}(\log n)^{\delta/\alpha}} \to 0 \qquad a.s.$$

for any $\delta > 0$.

REMARK. In the above example, for $0 < \alpha < 2$, $S_n/n^{1/\alpha}$ converges *in distribution* to a nondegenerate distribution called stable law. In particular, if $\alpha = 1$, S_n/n converges in distribution to a Cauchy distribution. For $\alpha = 2$, $S_n/(n \log n)^{1/2}$ converges to a normal distribution, and for $\alpha > 2$, $S_n/n^{1/2}$ converges to a normal distribution,

□

DIY EXERCISES.

Exercise 6.1. ⋆ ⋆ ⋆ Suppose $S_0 \equiv 0, S_1, S_2, ...$ form a square integrable martingale, i.e., for $k = 0, 1, ..., n$, $E(S_k^2) < \infty$ and $E(S_{k+1}|\mathcal{F}_k) = S_k$ where $\mathcal{F}_k$ is the σ-algebra generated by $S_1, ..., S_k$. Show that Kolmogorov's inequality still holds.

Exercise 6.2. ⋆ ⋆ ⋆ Prove the Corollary following Theorem 6.2.

Exercise 6.3. ⋆ ⋆ ⋆⋆ For positive independent r.v.s $X_1, X_2, ...$, show that the following three statements are equivalent: (a). $\sum_n X_n < \infty$ a.s.; (b). $\sum_n E(X_n \wedge 1) < \infty$; (c). $\sum_n E(X_n/(1 + X_n)) < \infty$.

Exercise 6.4. ⋆⋆⋆⋆ Raise a counterexample to show that there exists $X_1, X_2, ...$ iid with $E(X) = 0$ but $\sum_n X_n/n \not< \infty$ a.s..

Exercise 6.5 ⋆ ⋆ ⋆⋆ If $X_1, ...$ are iid with mean μ and finite variance. Then,

$$\frac{S_n - n\mu}{\sqrt{n(\log n)^\delta}} \to 0 \qquad a.s.$$

for any $\delta > 1$.

Exercise 6.6 ⋆⋆⋆ Suppose $X, X_1, ...$ are iid. Then, $(S_n - C_n)/n \to 0$ a.s. if and only if $E(|X|) < \infty$.

Exercise 6.7 ⋆ ⋆ ⋆⋆ Suppose $X, X_1, ...$ are iid with $E(|X|^p) = \infty$ for some $0 < p < \infty$. Then, $\limsup |S_n|/n^{1/p} = \infty$ a.s..

Exercise 6.8 ⋆ ⋆ ⋆⋆ Suppose $X_n, n \geq 1$ are independent with mean μ_n and variance σ_n^2 such that $\mu_n \to 0$ and $\sum_{j=1}^n \sigma_j^2 \to \infty$. show that

$$\frac{\sum_{j=1}^n X_j/\sigma_j^2}{\sum_{j=1}^n \sigma_j^{-2}} \to 0 \qquad a.s.$$

Hint: Consider the series $\sum_{j=1}^n (X_j - \mu_j)/(\sigma_j^2 \sum_{k=1}^j \sigma_k^{-2})$.

Chapter 7. Convergence in distribution and characteristic functions.

Convergence in distribution, which can be generalized slightly to weak convergence of measures, has been introduced in Chapter 3. This section provides a more detailed description.

(i). *Definition, basic properties and examples.*

Recall that in Section 1.3, we have already defined convergence in distribution for a sequence of random variables. Here we present the same definition in terms of weak convergence of their distributions. We first note that a function F is a cdf if and only if it is right continuous, nondecreasing with $F(t) \to 1$ and 0 when $t \to \infty$ and $-\infty$, respectively.

Definition. A sequence of distribution function F_n is called converging to another distribution function F_∞ *weakly*, if

(1) $F_n(t) \to F_\infty(t)$ for every continuity points of F_∞; or

(2), $\liminf_n F_n(B) \geq F_\infty(B)$ for every open set B in $(-\infty, \infty)$; or

(3) $\limsup_n F_n(C) \leq F_\infty(C)$ for every closed set C in $(-\infty, \infty)$; or

(4) $\int g(x)dF_n(x) \to \int g(x)dF_\infty(x)$ for every continuous function g.

Here $F_n(A)$ is defined as $\int_A dF_n(x) = \int 1_{x\in A} dF_n(x)$ for any Borel set A. The above four claims are equivalent to each other, as proved in Chapter 3.

Remark. If F_∞ is continuous, the inequalities in (2) and (3) are actually equalities. On the other hand, if X_n all takes integer values, then $X_n \to X$ in distribution is equivalent to $P(X_n = k) \to P(X = k)$ for all integer values k.

Remark. (Sheffe's Theorem) Suppose X_n has density function $f_n(\cdot)$ and $f_n(t) \to f(t)$ for every finite t and f *is a density function.* Then, $X_n \to X$ in distribution, where X has density f. This can be shown quite straightforwardly as follows:

$$\begin{aligned} 2 &= \int \liminf_n (f_n + f - |f_n(x) - f(x)|)dx \leq \liminf_n \int (f_n(x) + f(x) - |f_n(x) - f(x)|)dx \\ &= \liminf_n \Big(2 - \int |f_n(x) - f(x)|dx\Big) = 2 - \limsup_n \int |f_n(x) - f(x)|dx. \end{aligned}$$

Certainly, for any Borel set B,

$$P(X_n \in B) - P(X \in B) = \int_B (f_n(x) - f(x))dx \leq \int |f_n(x) - f(x)|dx \to 0.$$

□

In the above proof, we have used Fatou lemma with Lebesgue measure. In fact, the monotone convergence theorem, Fatou lemma and dominated convergence theorem that we have established with probability measure all hold with σ-finite measures, including Lebesgue measure.

Remark. (Slutsky's Theorem) Suppose $X_n \to X_\infty$ in distribution and $Y_n \to c$ in probability. Then, $X_n Y_n \to cX_\infty$ in distribution and $X_n + Y_n \to X_n - c$ in distribution.

We leave the proof as an exercise.

In the following, we provide some classical examples about convergence in distribution, only to show that there are a variety of important limiting distributions besides the normal distribution as the limiting distribution in CLT.

Example 7.1. (Convergence of maxima and extreme value distributions) Let $M_n = \max_{1\leq i\leq n} X_i$ where X_i are iid r.v.s with c.d.f. $F(\cdot)$. Then,

$$P(M_n \leq t) = P(X_1 \leq t)^n = F(t)^n.$$

As $n \to \infty$, the limiting distribution of properly scaled M_n, should it converge, should only be related with the right tail of the distribution of $F(\cdot)$, i.e., the $F(x)$ when x is large. The following are some examples.

(a). $F(x) = 1 - x^{-\alpha}$ for some $\alpha > 0$ and all large x. Then, for any $t > 0$,

$$P(M_n/n^{1/\alpha} < t) = (1 - n^{-1}t^{-\alpha})^n \to e^{-t^{-\alpha}}$$

(b). $F(x) = 1 - |x|^{\beta}$ for $x \in [-1, 0]$ and some $\beta > 0$. Then, for any $t < 0$,

$$P(n^{1/\beta}M_n \le t) = (1 - n^{-1}|t|^{\beta})^n \to e^{-|t|^{\beta}}$$

(c). $F(x) = 1 - e^{-x}$ for $x > 0$, i.e., X_i follows exponential distribution. Then for all t,

$$P(M_n - \log n \le t) \to e^{-e^{-t}}$$

These limiting distributions are called extreme value distributions.

EXAMPLE 7.2. (BIRTHDAY PROBLEM) Suppose $X_1, X_2, ...$ are iid with uniform distribution on the integers $\{1, 2, ..., N\}$ with $n < N$ and , Let

$$T_N = \min\{k : \text{ there exists a } j < k \text{ such that } \{X_j = X_k\} \}.$$

Then, for $k \le N$,

$$\begin{aligned} P(T_N > k) &= P(\ X_1, ..., X_k \text{ all take different values }) \\ &= \prod_{j=2}^{k} \Big(1 - P(\ X_j \text{ takes one of the values of } X_1, .., X_{j-1})\Big) \\ &= \prod_{j=2}^{k}(1 - \frac{j-1}{N}) = \exp\{\sum_{j=1}^{k-1}\log(1 - j/N)\} \end{aligned}$$

Then, for any fixed $x > 0$, as $N \to \infty$,

$$P(T_N/N^{1/2} > x) = P(T_N > N^{1/2}x) \approx \exp\{\sum_{1 \le j < N^{1/2}x} \log(1 - j/N)\}$$

$$\approx \exp\{-\sum_{1 \le j < N^{1/2}x} j/N\} \approx \exp\{-(1/N)N^{1/2}x(N^{1/2}x + 1)/2\} \approx \exp\{-x^2/2\}$$

In other words, $T_N/N^{1/2}$ converges in distribution to a distribution $F(t) = 1 - \exp(-t^2/2)$ for $t \ge 0$. Suppose now $N = 365$. By this approximation, we have $P(T_{365} > 22) \approx .5153$ and $P(T_{365} > 50) \approx .0326$, meaning that, with 22 (50) people there is about half (3%) probability that all of them have different birthday.

EXAMPLE 7.3. (LAW OF RARE EVENTS) Suppose there are totally n flights worldwide each year, and each flight has chance p_n to have an accident, independent of rest flights. There is on average λ accidents a year worldwide. The distribution of the number of accidents is $B(n, p_n)$ with np_n close to λ. Then this distribution approximates Poisson distribution with mean λ, namely,

$$Bin(n, p_n) \to \mathcal{P}(\lambda) \qquad \text{if} \quad n \to \infty \text{ and } np_n \to \lambda > 0.$$

Proof. For any fixed $k \ge 0$, and $n \ge k$

$$\begin{aligned} P(Bin(n, p_n) = k) &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{n!}{k!(n-k)!}\frac{(np_n)^k}{n^k}\frac{(1-p_n)^n}{(1-p_n)^k} \\ &= \frac{1}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{(np_n)^k e^{n\log(1-p_n)}}{(1-p_n)^k} \\ &\to \frac{\lambda^k e^{-\lambda}}{k!}, \qquad \text{as } n \to \infty. \end{aligned}$$

□

EXAMPLE 7.4. (THE SECRETARY/MARRIAGE PROBLEM) Suppose there are n secretary to be interviewed one by one and, right after each interview, you must make immediate decision of "hire or fire" the interviewee. You observe only the relative ranks of the interviewed candidates. What is the optimal strategy is maximize the chance of hiring the best of the n candidates? (Assume no ties of performance.)

One type of strategy is to give up the first m candidates, whatever their performance in the interview. Afterwards, the one that outperforms all previous candidates is hired. In other words, starting from $m+1$-th interview, the first candidate that outperforms the first m candidates is hired. Or else you settle with the last candidate. The chance that the k-th best among all n candidates is hired is

$$\begin{aligned} P_k &= \sum_{j=m+1}^{n} P(\text{ the } k\text{-th best is the } j\text{-th interviewee and is hired}) \\ &= \sum_{j=m+1}^{n} \frac{1}{n} P(\text{the best among first } j-1 \text{ appears in the first } m, \\ &\quad \text{the } j\text{-th candidate is the } k\text{-th best, and the } k-1 \text{ best all appear after the } j\text{-th candidate.}) \\ &\approx \sum_{j=m+1}^{n} \frac{m}{j-1} \times \frac{1}{n} \times (\frac{n-j}{n})^{k-1} \end{aligned}$$

Let $n \to \infty$, and $m \approx nc$ where c is the percentage of the interviews to be given up. Then the probability of hiring the k-th best

$$P_k \approx c \sum_{j=m}^{n} \frac{1}{j}(1-j/n)^{k-1} \approx c \int_c^1 \frac{(1-x)^{k-1}}{x} dx = cA_k, \quad \text{say.}$$

Since $A_{k+1} = A_k - (1-c)^k/k$, for $k \geq 1$, and $A_1 = -\log c$, it follows that

$$P_k \to c\Big(-\log c - \sum_{j=1}^{k-1} \frac{(1-c)^j}{j} \Big), \qquad \text{as} \qquad n \to \infty.$$

In particular, $P_1 \to -c \log c$. The function $c \log c$ is maximized at $c = 1/e = 0.368$. The best strategy is to give up the first 36.8% of the interviews and then hire the best to date. The chance of hiring the best overall is also 36.8%. The chance of hiring the last person is also c. This phenomenon is also called $1/e$ law. □

You might please formulate this problem in terms of a sequence of random variables.

(ii). *Some theoretical results about convergence in distribution.*

(a). FATOU LEMMA Suppose $X_n \geq 0$ and $X_n \to X_\infty$ in distribution. Then $E(X_\infty) \leq \liminf_n E(X_n)$.

Proof. Write

$$E(X_\infty) = \int_0^\infty P(X_\infty \geq t)dt \leq \int_0^\infty \liminf_n P(X_n \geq t)dt = \liminf_n \int_0^\infty P(X_n \geq t)dt \leq \liminf_n E(X_n).$$

□

The dominated convergence theorem also holds with convergence in distribution, which is left as an exercise.

(b). CONTINUOUS MAPPING THEOREM: $X_n \to X_\infty$ in distribution and $g(\cdot)$ is a continuous function. Then, $g(X_n) \to g(X_\infty)$ in distribution.

Proof. For any bounded continuous function f, $f(g(\cdot))$ is still bounded continuous function. Hence $E(f(g(X_n))) \to E(f(g(X_\infty)))$, proving that $g(X_n) \to g(X_\infty)$ in distribution. □

(c). Tightness and convergent subsequences.

In studying the convergence of a sequence of numbers, it is very useful that boundedness of the sequence, guarantees a convergent subsequence. The same is true for uniformly bounded monotone functions, such as, for example, distribution functions. This is the following Helly's Selection theorem, which is useful in studying weak convergence of distributions.

HELLY'S SELECTION THEOREM. A sequence of cumulative distribution functions F_n always contains a subsequence, say F_{n_k}, that converges to a function, say F_∞, which is nondecreasing and right continuous, at every continuity point of F_∞. If $F_\infty(-\infty) = 0$ and $F_\infty(\infty) = 1$. Then, F_∞ is a distribution function and F_{n_k} converges to F weakly.

Proof Let $t_1, t_2, ...$ be all rational numbers. In the sequence $F_n(t_1), n \geq 1$, there is always a convergent subsequence. Denote one of them as, say $n_k^{(1)}, k = 1, 2, ...$ Among this subsequence there is again a further subsequence, denoted as $n_k^{(2)}, k = 1, 2, ...,$ with $n_1^{(2)} > n_1^{(1)}$, such that $F_{n_k^{(2)}}(t_2)$ is convergent. Repeat this process of selection infinitely. Let $n_k = n_1^{(k)}$ be the first element of the k-th sub-sub-sequence. Then, for any fixed m, $\{n_k : k \geq m\}$ is always a subsequence of $\{n_k^{(l)} : k \geq 1\}$ for all $l \leq m$. Hence F_{n_k} is convergent on every rational number. Denote the limit as $F^*(t_l)$ on every rational t_l. Monotonicity of F_{n_k} implies the monotonicity of F^* on rational numbers. Define, for all t, $F_\infty(t) = \inf\{F^*(t_l) : t_l > t,\ t_l \text{ are rational}\}$. Than, F_∞ is right continuous and non-decreasing. The right continuity of F_n ensures that, if s is a continuity point of F_∞, $F_{n_k}(s) \to F_\infty(s)$. □

Not all sequence of distributions F_n would converge weakly to a *distribution function.* The easiest example is $F_n(\{n\}) = F_n(n) - F_n(n-) = 1$, i.e., $P(X_n = n) = 1$. Then, $F_n(t) \to 0$ for all $t \in (-\infty, \infty)$. If F_n all have little probability mass near ∞ or $-\infty$, then the convergence to a function which is not a distribution function can be avoided. A sequence of distribution functions F_n is called *tight* if, for any $\epsilon > 0$, there exists a $M > 0$ such that $\limsup_{n\to\infty}(1 - F_n(M) + F_n(M) < \epsilon$; Or, in other words,

$$\sup_n(1 - F_n(x) + F_n(-x)) \to 0 \qquad \text{as } x \to \infty.$$

PROPOSITION. Every tight sequence of distribution functions contains a a subsequence that weakly converges to a distribution function.

Proof Repeat the proof Helly's Selection Theorem. The tightness ensures the limit is a distribution function. □

(iii). *Characteristic functions.*

Characteristic function is one of the most useful tools in developing theory about convergence in distribution. The technical details of characteristic functions involve some knowledge of complex analysis. We shall view them as only a tool and try not to elaborate the technicalities.

1°. Definition and examples.

For a r.v. X with distribution F, its characteristic function is

$$\psi(t) = E(e^{itX}) = E(cos(tX) + isin(tX)) = \int e^{itx} dF(x), \qquad t \in (-\infty, \infty)$$

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

Some basic properties are:

$$\psi(0) = 1; \qquad |\psi(\cdot)| \leq 1; \qquad \psi(\cdot) \text{ is continuous on } (-\infty, \infty)$$

If ψ is characteristic function of X, then $e^{itb}\psi(at)$ is characteristic function of $aX + b$.

Product of characteristic functions is still a characteristic function. And the characteristic function of $X_1 + ... + X_n$ is the product of those of $X_1, ..., X_n$.

The following table lists some characteristic functions for some commonly used distributions:

Distribution	Density/Probability function	characteristic function (of t)
Degenerate	$P(X = a) = 1$	e^{iat}
Binomial $Bin(n, p)$	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, ..., n$	$(pe^{it} + 1 - p)^n$
Poisson $\mathcal{P}(\lambda)$:	$P(X = k) = \lambda^k e^{-\lambda}/k!$, $k = 0, 1, ...$	$\exp(\lambda(e^{it} - 1))$
Normal $N(\mu, \sigma^2)$:	$f(x) = e^{-(x-\mu)^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$, $x \in (-\infty, \infty)$	$e^{i\mu t - \sigma^2 t^2/2}$
Uniform $Unif[0, 1]$:	$f(x) = 1$, $x \in [0, 1]$	$(e^{it} - 1)/(it)$
Gamma :	$f(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x}/\Gamma(\alpha)$, $x > 0$	$(1 - it/\lambda)^{-\alpha}$
Cauchy:	$f(x) = 1/[\pi(1 + x^2)]$, $x \in (-\infty, \infty)$	$e^{-\lvert t \rvert}$

2°. Levy's inversion formula.

PROPOSITION Suppose X is r.v. with characteristic function $\psi(\cdot)$. Then, for all $a < b$,

$$\lim_{n\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \psi(t)dt = P(a < X < b) + \frac{1}{2}(P(X = a) + P(X = b)).$$

Proof. The proof uses Fubini's theorem to interchange the the expectation with the integration and the fact that $\int_0^\infty sin(x)/xdx = \pi/2$. We omit the proof.

The above theorem clearly implies that two different distribution cannot have same characteristic function, as formally presented in the following corollary.

Corollary. There is one-to-one correspondence between distribution functions and characteristic functions.

3°. Levy's continuity theorem.

Theorem 7.1 LEVY'S CONTINUITY THEOREM. *Let F_n, F_∞ be cdf with characteristic function ψ_n, ψ_∞. Then,*

(a). If $F_n \to F_\infty$ weakly, the $\psi_n(t) \to \psi(t)$ for every t.

(b). If $\psi_n(t) \to \psi(t)$ for every t, and $\psi(\cdot)$ is continuous at 0, then $F_n \to F$ weakly, where F is a cdf with characteristic function ψ.

Proof. Part (a) directly follows from the definition of convergence in distribution since e^{itx} is a continuous function of x for every t. Proof of part (b) uses the Levy inversion formula. We omit the details.

REMARK. Levy's continuity theorem enables us to show convergence of distribution through pointwise convergence of characteristic functions. This shall be our approach to establish the central limit theorem.

DIY EXERCISES:

Exercise 7.1. ⋆⋆⋆ Prove Slutsky's Theorem.

Exercise 7.2. ⋆⋆⋆ (DOMINATED CONVERGENCE THEOREM) Suppose $X_n \to X_\infty$ in distribution and $|X_n| \le Y$ with $E(Y) < \infty$. Show that $E(X_n) \to E(X_\infty)$.

Exercise 7.3. ⋆⋆ Suppose X_n is independent of Y_n, and X is independent of Y. Use characteristic functions to show that, if X_n converges to X in distribution and Y_n converges to Y in distribution and , then $X_n + Y_n$ converges in distribution to $X + Y$.

Chapter 8. Central limit theorem.

The most ideal case of the CLT is that the random variables are iid with finite variance. Although it is a special case of the more general Lindeberg-Feller CLT, it is most standard and its proof contains the essential ingredients to establish more general CLT. Throughout the chapter, $\Phi(\cdot)$ is the cdf of standard normal distribution $N(0,1)$.

(i). Central limit theorem (CLT) for iid r.v.s.

The following lemma plays a key role in the proof of CLT.

Lemma 8.1 For any real x and $n \geq 1$,

$$|e^{ix} - \sum_{j=0}^{n} \frac{(ix)^j}{j!}| \leq \min\Big(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\Big).$$

Consequently, for any r.v. X with characteristic function ψ and finite second moment,

$$\Big|\psi(t) - [1 + itE(X) - \frac{t^2}{2}E(X^2)]\Big| \leq \frac{|t|^2}{6}E(\min(|t||X|^3,\ 6|X|^2)). \tag{8.1}$$

Proof. The proof relies on the identity

$$e^{ix} - \sum_{j=0}^{n} \frac{(ix)^j}{j!} = \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds = \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1)ds,$$

which can be shown by induction and by taking derivatives. The middle term is bounded by $|x|^{n+1}/(n+1)!$, and the last bounded by $2|x|^n/n!$. □

Theorem 8.2 *Suppose $X, X_1, ..., X_n, ...$ are iid with mean μ and finite variance $\sigma^2 > 0$. Then,*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \to N(0,1) \qquad \textit{in distribution.}$$

Proof. Without loss of generality, let $\mu = 0$. Let ψ be the common characteristic function of X_i. Observe that, by dominated convergence

$$E(\min(|t_n||X|^3,\ 6|X|^2)) \to 0 \quad \text{as} \quad |t_n| \to 0$$

The characteristic function of $S_n/\sqrt{n\sigma^2}$ is, by applying the above lemma,

$$\begin{aligned} E(e^{itS_n/\sqrt{n\sigma^2}}) &= E(e^{it_n S_n}) = \prod_{j=1}^{n} E(e^{itX_j/\sqrt{n\sigma^2}}) = \psi^n(\frac{t}{\sqrt{n\sigma^2}}) \\ &= [1 + \frac{it}{\sqrt{n\sigma^2}}E(X) - \frac{t^2}{2n\sigma^2}E(X^2) + o(\frac{1}{n})]^n = [1 - \frac{t^2}{2n} + o(\frac{1}{n})]^n \\ &\to e^{-t^2/2}, \end{aligned}$$

which is the characteristic function of $N(0,1)$. Then, Levy's continuity theorem implies the above CLT. □

In the case the common variance is not finite, the partial sum, after proper normalization, may or may not converge to a normal distribution. The following theorem provides sufficient and necessary condition. The key point here is whether there exists appropriate truncation, which is a trick that we have used so many times before.

Theorem 8.3 *Suppose $X, X_1, X_2, ...$ are iid nondegenerate. Then, $(S_n - a_n)/b_n$ converges to a normal distribution for some constants a_n and $0 < b_n \to \infty$, if and only if*

$$\frac{x^2 P(|X| > x)}{E(X^2 1_{\{|X| \le x\}})} \to 0, \qquad \text{as } x \to \infty. \tag{8.2}$$

The proof is omitted. We note that (8.2) holds if X_i has finite variance $\sigma^2 > 0$, in which case CLT of Theorem 8.2 holds with $a_n = nE(X)$ and $b_n = \sqrt{n}\sigma$. Theorem 8.3 is of interest when $E(X^2) = \infty$. In this case, one can choose to truncate the X_is at

$$c_n = \sup\{c : nE(|X|^2 1_{\{|X| \le c\}})/c^2 \ge 1\}$$

With some calculation, condition (8.2) ensures

$$nP(|X| > c_n) \to 0 \qquad \text{and} \qquad nE(|X|^2 1_{\{|X| \le c_n\}})/c_n^2 \to 1.$$

Separate S_n into two parts, one with X_i beyond $\pm c_n$ and the other bounded by $\pm c_n$. The former takes value 0 with chance going to 1. The latter, when standardized by

$$a_n = nE(X 1_{\{|X| \le c_n\}}) \qquad \text{and} \qquad b_n = \sqrt{nE(X^2 1_{\{|X| \le c_n\}})} \approx c_n.$$

converges to $N(0,1)$, which can be shown by repeating the proof of Theorem 8.2 or by citing Lindeberg-Feller CLT. We note that $b_n \approx \sqrt{n\text{var}(X 1_{\{|X| \le c_n\}})}$ by (8.2).

EXAMPLE 8.1 Recall Example 6.3, in which $X, X_1, X_2, ...$ are iid symmetric such that $P(|X| > x) = x^{-\alpha}$ for some $\alpha > 0$ all large x. Then, Theorem 8.3 implies $(S_n - a_n)/b_n \to N(0,1)$ if and only if $\alpha \ge 2$. Indeed, when $\alpha > 2$, the common variance is finite and CLT applies. When $\alpha = 2$,

$$S_n/(n \log n)^{1/2} \to N(0, \sigma^2)$$

for some σ^2.

When $\alpha < 2$, the condition in Theorem cannot hold. In fact, S_n when properly normalized shall converge to non-normal distribution.

(ii). *The Lindeberg-Feller CLT.*

Theorem 8.4 LINDEBERG-FELLER CLT. *Suppose $X_1, ..., X_n, ...$ are independent r.v.s with mean 0 and variance σ_n^2. Let $s_n^2 = \sum_{j=1}^n \sigma_j^2$ denote the variance of partial sum $S_n = X_1 + \cdots + X_n$. If, for every $\epsilon > 0$,*

$$\frac{1}{s_n^2} \sum_{j=1}^n E(X_j^2 1_{\{|X_j| > \epsilon s_n\}}) \to 0, \tag{2.3}$$

then $S_n/s_n \to N(0,1)$. Conversely, if $\max_{j \le n} \sigma_j^2/s_n^2 \to 0$ and $S_n/s_n \to N(0,1)$, then (8.3) holds.

Proof. "$\Longleftarrow$" The Lindeberg condition (8.3) implies

$$\max_{1 \le j \le n} \left(\frac{\sigma_j^2}{s_n^2}\right) \le \epsilon^2 + \frac{1}{s_n^2} \max_{1 \le j \le n} E(X_j^2 1_{\{|X_j| > \epsilon s_n\}}) \to 0, \tag{8.4}$$

by letting $n \to \infty$ and then $\epsilon \downarrow 0$. Observe that for every real $x > 0$, $|e^{-x} - 1 + x| \le x^2/2$. Moreover, for complex z_j and w_j with $|z_j| \le 1$ and $|w_j| \le 1$,

$$|\prod_{j=1}^n z_j - \prod_{j=1}^n w_j| \le \sum_{j=1}^n |z_j - w_j|, \tag{8.5}$$

which can be proved by induction. With Lemma 2.1, it follows that, for any $\epsilon > 0$,

$$\begin{aligned}
& |E(e^{itX_j/s_n}) - e^{-t^2\sigma_j^2/2s_n^2}| \\
\leq\ & |E\Big(1 + itX_j - \frac{(tX_j)^2}{2s_n^2}\Big) - \Big(1 - \frac{t^2\sigma_j^2}{2s_n^2}\Big)| + E\Big[\min\Big(\frac{t^2X_j^2}{s_n^2}, \frac{|tX_j|^3}{6s_n^3}\Big)\Big] + \frac{t^4\sigma_j^4}{8s_n^4} \\
\leq\ & E\Big(\frac{t^2X_j^2}{s_n^2}1_{\{|X_j|>\epsilon s_n\}}\Big) + E\Big(\frac{|tX_j|^3}{6s_n^3}1_{\{|X_j|\leq\epsilon s_n\}}\Big) + \frac{t^4\sigma_j^4}{8s_n^4} \\
\leq\ & \frac{t^2}{s_n^2}E(X_j^2 1_{\{|X_j|>\epsilon s_n\}}) + \frac{|t|^3\epsilon}{s_n^2}E(X_j^2) + \frac{t^4\sigma_j^2}{s_n^2}\max_{1\leq k\leq n}\frac{\sigma_k^2}{s_n^2}
\end{aligned}$$

Then, for any fixed t,

$$\begin{aligned}
& |E(e^{itS_n/s_n}) - e^{-t^2/2}| \\
=\ & |\prod_{j=1}^n E(e^{itX_j/s_n}) - \prod_{j=1}^n e^{-t^2\sigma_j^2/2s_n^2}| \\
\leq\ & \sum_{j=1}^n |E(e^{itX_j/s_n}) - e^{-t^2\sigma_j^2/2s_n^2}| \qquad \text{by (8.5)} \\
\leq\ & \sum_{j=1}^n \Big(\frac{t^2}{s_n^2}E(X_j^2 1_{\{|X_j|>\epsilon s_n\}}) + \frac{|t|^3\epsilon}{s_n^2}E(X_j^2) + \frac{t^4\sigma_j^2}{s_n^2}\max_{1\leq j\leq n}\frac{\sigma_j^2}{s_n^2}\Big) \\
\leq\ & \Big(\frac{t^2}{s_n^2}\sum_{j=1}^n E(X_j^2 1_{\{|X_j|>\epsilon s_n\}}) + \epsilon|t|^3 + t^4\max_{1\leq j\leq n}\frac{\sigma_j^2}{s_n^2}\Big) \\
\to\ & \epsilon|t|^3, \qquad \text{as } n\to\infty, \qquad \text{by (8.3) and (8.4).}
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, it follows that $E(e^{itS_n/s_n}) \to e^{-t^2/2}$ for all t. Levy's continuity theorem implies $S_n/s_n \to N(0,1)$.

"$\Longleftarrow$" Let ψ_j be the moment generating function of X_j. The asymptotic normality is equivalent to $\prod_{j=1}^n \psi_j(t/s_n) \to e^{-t^2/2}$. Notice that (8.1) implies

$$|\psi_j(t/s_n) - 1| \leq 2\frac{t^2\sigma_j^2}{s_n} \tag{8.6}$$

Write, as $n\to\infty$,

$$\begin{aligned}
& \sum_{j=1}^n [\psi_j(t/s_n) - 1] + t^2/2 \\
=\ & \sum_{j=1}^n [\psi_j(t/s_n) - 1 - \log\psi_j(t/s_n)] + \sum_{j=1}^n [\log\psi_j(t/s_n)] + t^2/2 \\
\leq\ & \sum_{j=1}^n |\psi_j(t/s_n) - 1 - \log\psi_j(t/s_n)| + +o(1) \\
\leq\ & \sum_{j=1}^n |\psi_j(t/s_n) - 1|^2 + o(1) \\
\leq\ & \max_{1\leq k\leq n}|\psi_k(t/s_n) - 1| \times \sum_{j=1}^n |\psi_j(t/s_n) - 1| + o(1) \\
\leq\ & 4\max_{1\leq k\leq n}\frac{t^2\sigma_k^2}{s_n} \times \sum_{j=1}^n \frac{t^2\sigma_j^2}{s_n} + o(1) \qquad \text{by (8.6)} \\
=\ & o(1), \qquad \text{by the assumption } \max_{j\leq n}\sigma_j^2/s_n^2 \to 0.
\end{aligned}$$

On the other hand, by definition of characteristic function, the above expression is, as $n \to \infty$,

$$\begin{aligned} o(1) &= \sum_{j=1}^{n} [\psi_j(t/s_n) - 1] + t^2/2 \\ &= \sum_{j=1}^{n} E(e^{itX_j/s_n} - 1) + t^2/2 = \sum_{j=1}^{n} E(cos(tX_j/s_n) - 1) + t^2/2 + i\sum_{j=1}^{n} E(sin(tX_j/s_n)) \\ &= \sum_{j=1}^{n} E\{(cos(tX_j/s_n) - 1)1_{\{|X_j|>\epsilon s_n\}}\} + \sum_{j=1}^{n} E\{(cos(tX_j/s_n) - 1)1_{\{|X_j|\le\epsilon s_n\}}\} + t^2/2 \\ &\qquad +\text{imaginary part (immaterial).} \end{aligned}$$

Since $cos(x) - 1 \ge -x^2/2$ for all real x,

$$\begin{aligned} \frac{1}{s_n^2}\sum_{j=1}^{n} E(X_j^2 1_{\{X_j|>\epsilon s_n\}}) &= 1 - \frac{2}{t^2}\sum_{j=1}^{n} E(\frac{t^2X_j^2}{2s_n^2}1_{\{X_j|\le\epsilon s_n\}}) \\ &\le \frac{2}{t^2}\Big(\frac{t^2}{2} + \sum_{j=1}^{n} E\{(cos(tX_j/s_n) - 1)1_{\{|X_j|\le\epsilon s_n\}}\}\Big) \\ &\le \frac{2}{t^2}\Big(|\sum_{j=1}^{n} E\{(cos(tX_j/s_n) - 1)1_{\{|X_j|>\epsilon s_n\}}\}| + o(1)\Big) \\ &\le \frac{2}{t^2}\sum_{j=1}^{n} 2P(|X_j| > \epsilon s_n) + o(1) \\ &\le \frac{4}{t^2}\sum_{j=1}^{n} \frac{\sigma_j^2}{(\epsilon s_n)^2} + o(1) \qquad \text{by Chebyshev inequality} \\ &\le \frac{4}{t^2\epsilon^2} + o(1). \end{aligned}$$

Since t can be chosen arbitrarily large, Lindeberg condition holds. □

Remark. Sufficiency is proved by Lindeberg in 1922 and necessity by Feller in 1935. Lindeberg-Feller CLT is one of the most far-reaching results in probability theory. Nearly all generalizations of various types of central limit theorems spin from Lindeberg-Feller CLT, such as, for example, CLT for martingales, for renewal proceses, or for weakly dependent processes. The insights of the Lindeberg condition (8.3) are that the "wild" values of the random variables, compared with s_n, the standard deviation of S_n as the normalizing constant, are insignificant and can be truncated off without affecting the general behavior of the partial sum S_n.

Example 8.2. Suppose X_n are independent and

$$P(X_n = n) = P(X_n = -n) = n^{-\alpha}/4 \qquad \text{and} \qquad P(X_n = 0) = 1 - n^{-\alpha}/2,$$

with $0 < \alpha < 3$. Then, $\sigma_n^2 = E(X_n^2) = n^{2-\alpha}/2$ and $s_n^2 = \sum_{j=1}^{n} j^{2-\alpha}/2$, which increases to ∞ at the order of $n^{3-\alpha}$. Note that Lindeberg condition (2.3) is equivalent to $n^2/n^{3-\alpha} \to 0$, i.e., $0 < \alpha < 1$. On the other hand, $\max_{1\le j\le n} \sigma_j^2/s_n^2 \to 0$. Therefore, it follows from Theorem 8.4 that $S_n/s_n \to N(0,1)$ if and only if $0 < \alpha < 1$. □

Example 8.3 Suppose X_n are independent and $P(X_n = 1) = 1/n = 1 - P(X_n = 0)$. Then,

$$[S_n - \log(n)]/\sqrt{\log(n)} \to N(0,1) \qquad \text{in distribution.}$$

It's clear that $E(X_n) = 1/n$ and $\text{var}(X_n) = (1 - 1/n)/n$. So, $E(S_n) = \sum_{i=1}^{n} = \sum_{i=1}^{n} 1/i$, and $\text{var}(S_n) = \sum_{i=1}^{n}(1 - 1/i)/i \approx \log(n)$. As X_n are all bounded by 1 and $\text{var}(S_n) \uparrow \infty$, the Lindeberg

condition is satisfied. Therefore, by the CLT,

$$\frac{S_n - \sum_{i=1}^n 1/i}{[\sum_{i=1}^n (1-1/i)/i]^{1/2}} \to N(0,1), \qquad \text{in distribution.}$$

Then, $[S_n - \log(n)]/\sqrt{\log(n)} \to N(0,1)$ in distribution since $|\log(n) - \sum_{i=1}^n 1/i| \le 1$ and $\text{var}(S_n)/\log(n) \to 1$. □

Theorem 8.2 as well as the following Lyapunov CLT are both special cases of the Lindeberg-Feller CLT. Nevertheles they are convenient for application.

Corollary (LYAPUNOV CLT) *Suppose X_n are indendent with mean 0 and $\sum_{j=1}^n E(|X_j|^\delta)/s_n^\delta \to 0$ for some $\delta > 2$, then $S_n/s_n \to N(0,1)$.*

Proof. For any $\epsilon > 0$, as $n \to \infty$,

$$\frac{1}{s_n^2}\sum_{j=1}^n E(X_j^2 1_{\{|X_j| > \epsilon s_n\}}) = \sum_{j=1}^n E(\frac{X_j^2}{s_n^2} 1_{\{|X_j|/s_n > \epsilon\}}) \le \frac{1}{\epsilon^{\delta-2}} \sum_{j=1}^n E(\frac{X_j^\delta}{s_n^\delta}) \to 0.$$

Lindeberg condition (8.3) holds and hence CLT holds. □

In Example 8.2, for any $\delta > 2$, $\sum_{j=1}^n E|X_j|^\delta = \sum_{j=1}^n j^\delta j^{-\alpha}/2$ which increasing at the order $n^{\delta-\alpha+1}$, while s_n^δ increases at the order of $n^{(3-\alpha)\delta/2}$. Simple calculation shows, when $0 < \alpha < 1$, Lypunov CLT holds.

(iii). *CLT for arrays of random variables.*

Very often Lindeberg-Feller CLT is presented in the form of arrays of random variables as given in the textbook.

Theorem 8.5 (CLT FOR ARRAYS OF R.V.S) *Let $X_{n,1}, ..., X_{n,n}$ be n independent random variables with mean 0 such that, as $n \to \infty$,*

$$\sum_{j=1}^n \text{var}(X_{n,j}) \to 1 \qquad \textit{and} \qquad \sum_{j=1}^n E(X_{n,j}^2 1_{\{|X_{n,j}| > \epsilon\}}) \to 0, \qquad \textit{for any } \epsilon > 0.$$

Then, $S_n \equiv X_{n,1} + \cdots + X_{n,n} \to N(0,1)$.

This theorem is slightly more general than Lindeberg-Feller CLT, although the proof is identical to that of the first part of Theorem 8.4. Theorem 8.4. is a special case of Theorem 8.5 by letting $X_{n,i} = X_i/s_n$. Thus $X_{n,k}$ are undertood as the usual r.v.s normalized by the standard deviation of the partial sums. Thus S_n in this theorem is already standardized.

DIY EXERCISES

Exercise 8.1 ⋆⋆⋆ Suppse X_n are independent with

$$P(X_n = n^\alpha) = P(X_n = -n^\alpha) = \frac{1}{2n^\beta} \qquad \text{and} \qquad P(X_n = 0) = 1 - \frac{1}{n^\beta}$$

with $2\alpha > \beta - 1$. Show that the Lindeberg condition holds if and only if $0 \le \beta < 1$.

Exercise 8.2 ⋆⋆⋆ Suppse X_n are iid with mean 0 and variance 1. Let $a_n > 0$ be such that $s_n^2 = \sum_{j=1}^n a_i^2 \to \infty$ and $a_n/s_n \to 0$. Show that $\sum_{i=1}^n a_i X_i/s_n \to N(0,1)$.

Exercise 8.3 ⋆⋆⋆ Suppse $X_1, X_2...$ are independent and $X_n = Y_n + Z_n$, where Y_n takes values 1 and -1 with chance 1/2 each, and $P(Z_n = \pm n) = 1/(2n^2) = (1 - P(Z_n = 0))/2$. Show that Lindeberg condition does not hold, yet $S_n/\sqrt{n} \to N(0,1)$.

Exercise 8.4 ⋆⋆⋆ Suppse $X_1, X_2, ...$ are iid nonnegative r.v.s with mean 1 and finite variance $\sigma^2 > 0$. Show that $2(\sqrt{S_n} - \sqrt{n}) \to N(0,1)$.

.

Review of Probability Theory

1. Probability Calculation.

Calculation of probabilities of events for discrete outcomes (e.g., coin tossing, roll of dice, etc.)

Calculation of the probability of certain events for given density functions of 1 or 2 dimension.

2. Probability space.

(1). Set operations: $\cup$, $\cap$, complement.

(2). σ-algebra. Definition and implications (the collection of sets which is *closed* on set operations).

(3). Kolmogorov's trio of probability space.

(4). Independence of events and conditional probabilities of events.

(5). Borel-Cantelli lemma.

(6). Sets and set-index functions. $1_{A\cap B} = 1_A 1_B$ and $1_{A^c} = 1 - 1_A$. $1_{\{A_n, i.o.\}} = \limsup_n 1_{A_n}$. etc.

$$\{A_n, i.o.\} = \cap_{n=1}^{\infty} \cup_{k=1}^{\infty} A_k = \lim_{n\to\infty} \cup_{k=n}^{\infty} A_k = \limsup_n A_n \qquad (\limsup_n 1_{A_n}(\omega)).$$

$\omega \in \{A_n, i.o.\}$ means $\omega \in A_n$ for infinitely many A_n. (Mathematically precisely, there exists a subsequence $n_k \to \infty$, such that $\omega \in A_{n_k}$ for all n_k.)

$$\cup_{n=1}^{\infty} \cap_{k=1}^{\infty} A_k = \lim_{n\to\infty} \cap_{k=n}^{\infty} A_k = \liminf_n A_n \qquad (\liminf_n 1_{A_n}(\omega)).$$

$\omega \in \liminf_n A_n$ means $\omega \in A_n$ for all large n. (Mathematically precisely, there exists an N, such that $\omega \in A_n$ for all $n \geq N$.)

(7). $\star$ completion of probability space.

3. Random variables.

(1). Definitions.

(2). c.d.f., density or probability functions.

(3). Expectation: definition, interpretation as weighted (by chance) average.

(4). Properties:

(i). Dominated convergence. (Proof and Application).

(ii). $\star$ Fatou's lemma and monotone convergence.

(iii). Jensen's inequalities

(iv) Chebyshev inequalities.

(5). Independence of r.v.s.

(6). $\star$ Conditional distribution and expectation given a σ-algebra (Definition and simple properties.)

(7). Commonly used distributions and r.v.s.

4. Convergence.

(1). Definition of a.e., in prob, L^p and in distribution convergence.

(2). $\star$ Equivalence of four definitions of in distribution convergence.

(3). The diagram about the relation among convergence modes.

(4). The technique of truncating off r.v.s

5. LLN.

(1). WLLN. Application of the theorem. ($\star$ the proof)

(2). Kolmogorov's inequality ($\star$ the proof).

(3). Corollary ($\star$ the proof)

(4). $\star$ Kolmogorov's 3-series theorem

(5). Kolmogorov's criterion for SLLN. ($\star$ the proof)

(6) Kronecker lemma. ($\star$ the proof)

(7). SLLN for iid r.v.s ($\star$ the proof)

(8). Application.

6. CLT.

(1) Characteristic function. (Definition and simple properties.)

(i) ψ continuous. $|\psi| \leq 1$ and $\psi(0) = 1$.

(ii) $F_X = F_Y \Longrightarrow \psi_X = \psi_Y$. ($\star$ proof of $\Longleftarrow$.)

(iii) $F_n \to F \Longrightarrow \psi_n \to \psi$. ($\star$ proof of $\Longleftarrow$.)

(iv) X and Y are independent $\Longrightarrow \psi_{X+Y} = \psi_X + \psi_Y$.

(2). CLT for iid r.v.s

Theorem, application and the proof for the case of bounded r.v.s.

(3). Lindeberg condition.

Application and Heuristic interpretation. ($\star$ proof.)

(4). Application.

Remark. $\star$ means not required in the midterm exam.