Chapter 5

Elements of Probability Theory

The purpose of this chapter is to summarize some important concepts and results in probability theory. Of particular interest to us are the limit theorems which are powerful tools to analyze the convergence behaviors of econometric estimators and test statistics. These properties are the core of the asymptotic analysis in subsequent chapters. For a more complete and thorough treatment of probability theory see Davidson (1994) and other probability textbooks, such as Ash (1972) and Billingsley (1979). Bierens (1994), Gallant (1997) and White (2001) also provide concise coverages of the topics in this chapter. Many results here are taken freely from the references cited above; we will not refer to them again in the text unless it is necessary.

5.1 Probability Space and Random Variables

5.1.1 Probability Space

The probability space associated with a random experiment is determined by three components: the *outcome space* Ω whose element ω is an *outcome* of the experiment, a collection of events $\mathcal{F}$ whose elements are subsets of Ω, and a *probability measure* $\mathbb{P}$ assigned to the elements in $\mathcal{F}$.

Given the subset A of Ω, its complement is defined as $A^c = \{\omega \in \Omega : \omega \notin A\}$. In the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{F}$ is a σ-algebra (σ-field) in the sense that it satisfies the following requirements.

1. $\Omega \in \mathcal{F}$.

2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

3. If $A_1, A_2, \ldots$ are in $\mathcal{F}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

The first and second properties together imply that $\Omega^c = \emptyset$ is also in $\mathcal{F}$. Combining the second and third properties we have from de Morgan's law that

$$\left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} A_n^c \in \mathcal{F}.$$

A σ-algebra is thus closed under complementation, countable union and countable intersection.

The probability measure $\mathbb{P} : \mathcal{F} \mapsto [0,1]$ is a real-valued set function satisfying the following axioms.

1. $\mathbb{P}(\Omega) = 1$.
2. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.
3. if $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

From these axioms we easily deduce that $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, $\mathbb{P}(A) \leq \mathbb{P}(B)$ if $A \subseteq B$, and

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

Moreover, if $\{A_n\}$ is an increasing (decreasing) sequence in $\mathcal{F}$ with the limiting set A, then $\lim_n \mathbb{P}(A_n) = \mathbb{P}(A)$.

Let $\mathcal{C}$ be a collection of subsets of Ω. The intersection of all the σ-algebras that contain $\mathcal{C}$ is the smallest σ-algebra containing $\mathcal{C}$; see Exercise 5.1. This σ-algebra is referred to as the σ-algebra generated by $\mathcal{C}$, denoted as $\sigma(\mathcal{C})$. When $\Omega = \mathbb{R}$, the *Borel field* is the σ-algebra generated by all open intervals (a, b) in $\mathbb{R}$, usually denoted as $\mathcal{B}^d$. Note that open intervals, closed intervals $[a, b]$, half-open intervals $(a, b]$ or half lines $(-\infty, b]$ can be obtained from each other by taking complement, union and/or intersection. For example,

$$(a, b] = \bigcap_{n=1}^{\infty} \Big(a,\ b + \frac{1}{n}\Big), \qquad (a, b) = \bigcup_{n=1}^{\infty} \Big(a, b - \frac{1}{n}\Big].$$

Thus, the collection of all closed intervals (half-open intervals, half lines) generates the same Borel field. This is why open intervals, closed intervals, half-open intervals and half lines are also known as *Borel sets*. The Borel field on $\mathbb{R}^d$, denoted as $\mathcal{B}^d$, is generated by all open hypercubes:

$$(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d).$$

Equivalently, $\mathcal{B}^d$ can be generated by all closed hypercubes:

$$[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d],$$

or by

$$(-\infty, b_1] \times (-\infty, b_2] \times \cdots \times (-\infty, b_d].$$

The sets that generate the Borel field $\mathcal{B}^d$ are all Borel sets.

5.1.2 Random Variables

A random variable z defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $z\colon \Omega \mapsto \mathbb{R}$ such that for every B in the Borel field $\mathcal{B}$, its *inverse image* of B is in $\mathcal{F}$, i.e.,

$$z^{-1}(B) = \{\omega\colon z(\omega) \in B\} \in \mathcal{F}.$$

We also say that z is a $\mathcal{F}/\mathcal{B}$-measurable (or simply $\mathcal{F}$-measurable) function. Non-measurable functions are very exceptional in practice and hence are not of general interest. Given the random outcome ω, the resulting value $z(\omega)$ is known as a *realization* of z. The realization of z varies with ω and hence is governed by the random mechanism of the underlying experiment.

A $\mathbb{R}^d$-valued random variable (random vector) $\boldsymbol{z}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $\boldsymbol{z}\colon \Omega \mapsto \mathbb{R}^d$ such that for every $B \in \mathcal{B}^d$,

$$\boldsymbol{z}^{-1}(B) = \{\omega\colon \boldsymbol{z}(\omega) \in B\} \in \mathcal{F};$$

that is, $\boldsymbol{z}$ is a $\mathcal{F}/\mathcal{B}^d$-measurable function. Given the random vector $\boldsymbol{z}$, its inverse images $\boldsymbol{z}^{-1}(B)$ form a σ-algebra, denoted as $\sigma(\boldsymbol{z})$. This σ-algebra must be in $\mathcal{F}$, and it is the smallest σ-algebra contained in $\mathcal{F}$ such that $\boldsymbol{z}$ is measurable. This is known as the σ-algebra generated by $\boldsymbol{z}$ or, more intuitively, the information set associated with $\boldsymbol{z}$.

A function $g\colon \mathbb{R} \mapsto \mathbb{R}$ is said to be $\mathcal{B}$-measurable or *Borel measurable* if

$$\{\zeta \in \mathbb{R}\colon g(\zeta) \leq b\} \in \mathcal{B}.$$

If z is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$, then $g(z)$ is also a random variable defined on the same probability space provided that g is Borel measurable. Note that the functions we usually encounter (e.g., continuous functions and integrable functions) are Borel measurable. Similarly, for the d-dimensional random vector $\boldsymbol{z}$, $g(\boldsymbol{z})$ is a random variable provided that g is $\mathcal{B}^d$-measurable.

Recall from Section 2.1 that the joint distribution function of $\boldsymbol{z}$ is the non-decreasing, right-continuous function $F_{\boldsymbol{z}}$ such that for $\boldsymbol{\zeta} = (\zeta_1 \ \ldots \ \zeta_d)' \in \mathbb{R}^d$,

$$F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \mathbb{P}\{\omega \in \Omega\colon z_1(\omega) \leq \zeta_1, \ldots, z_d(\omega) \leq \zeta_d\},$$

with

$$\lim_{\zeta_1 \to -\infty, \ldots, \zeta_d \to -\infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 0, \qquad \lim_{\zeta_1 \to \infty, \ldots, \zeta_d \to \infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 1.$$

The marginal distribution function of the i^{th} component of $\boldsymbol{z}$ is such that

$$F_{z_i}(\zeta_i) = \mathbb{P}\{\omega \in \Omega \colon z_i(\omega) \leq \zeta_i\} = F_{\boldsymbol{z}}(\infty, \ldots, \infty, \zeta_i, \infty, \ldots, \infty).$$

Note that while $\mathbb{P}$ is a set function defined on $\mathcal{F}$, the distribution function of $\boldsymbol{z}$ is a point function defined on $\mathbb{R}^d$.

Two random variables y and z are said to be (pairwise) *independent* if, and only if, for any Borel sets B_1 and B_2,

$$\mathbb{P}(y \in B_1 \text{ and } z \in B_2) = \mathbb{P}(y \in B_1)\, \mathbb{P}(z \in B_2).$$

This immediately leads to the standard definition of independence: y and z are independent if, and only if, their joint distribution is the product of their marginal distributions, as in Section 2.1. A sequence of random variables $\{z_i\}$ is said to be *totally independent* if

$$\mathbb{P}\left(\bigcap_{\text{all } i} \{z_i \in B_i\}\right) = \prod_{\text{all } i} \mathbb{P}(z_i \in B_i),$$

for any Borel sets B_i. In what follows, a totally independent sequence will be referred to an independent sequence or a sequence of independent variables for convenience. For an independent sequence, we have the following generalization of Lemma 2.1.

Lemma 5.1 *Let $\{z_i\}$ be a sequence of independent random variables and h_i, $i = 1, 2, \ldots$, be Borel-measurable functions. Then $\{h_i(z_i)\}$ is also a sequence of independent random variables.*

5.1.3 Moments and Norms

The expectation of the i^{th} element of $\boldsymbol{z}$ is

$$\mathbb{E}(z_i) = \int_\Omega z_i(\omega)\, \mathrm{d}\, \mathbb{P}(\omega),$$

where the right-hand side is a Lebesgue integral. In view of the distribution function defined above, a change of ω causes the realization of $\boldsymbol{z}$ to change so that

$$\mathbb{E}(z_i) = \int_{\mathbb{R}^d} \zeta_i \, \mathrm{d} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \int_{\mathbb{R}} \zeta_i \, \mathrm{d} F_{z_i}(\zeta_i),$$

where F_{z_i} is the marginal distribution function of the i^{th} component of $\boldsymbol{z}$, as defined in Section 2.2. For the Borel measurable function g of $\boldsymbol{z}$,

$$\mathbb{E}[g(\boldsymbol{z})] = \int_\Omega g(\boldsymbol{z}(\omega))\, \mathrm{d}\, \mathbb{P}(\omega) = \int_{\mathbb{R}^d} g(\boldsymbol{\zeta}) \, \mathrm{d} F_{\boldsymbol{z}}(\boldsymbol{\zeta}).$$

Other moments, such as variance and covariance, can also be defined as Lebesgue integrals with respect to the probability measure; see Section 2.2.

A function g is said to be *convex* on a set S if for any $a \in [0,1]$ and any x, y in S,

$$g\big(ax + (1-a)y\big) \leq ag(x) + (1-a)g(y);$$

g is *concave* on S if the inequality above is reversed. For example, $g(x) = x^2$ is convex, and $g(x) = \log x$ for $x > 0$ is concave. The result below is concerned with convex (concave) transformations of random variables.

Lemma 5.2 (Jensen) *For the Borel measurable function g that is convex on the support of the integrable random variable z, suppose that $g(z)$ is also integrable. Then,*

$$g(\mathbb{E}(z)) \leq \mathbb{E}[g(z)];$$

the inequality reverses if g is concave.

For the random variable z with finite p^{th} moment, let $\|z\|_p = [\mathbb{E}(z^p)]^{1/p}$ denote its L_p-norm. Also define the *inner product* of two square integrable random variables z_i and z_j as their cross moment:

$$\langle z_i, z_j \rangle = \mathbb{E}(z_i z_j).$$

Then, L_2-norm can be obtained from the inner product as $\|z_i\|_2 = \langle z_i, z_i \rangle^{1/2}$. It is easily seen that for any $c > 0$ and $p > 0$,

$$c^p\, \mathbb{P}(|z| \geq c) = c^p \int \mathbf{1}_{\{\zeta : |\zeta| \geq c\}}\, \mathrm{d}F_z(\zeta) \leq \int_{\{\zeta : |\zeta| \geq c\}} |\zeta|^p\, \mathrm{d}F_z(\zeta) \leq \mathbb{E}\,|z|^p,$$

where $\mathbf{1}_{\{\zeta : |\zeta| \geq c\}}$ is the indicator function which equals one if $|\zeta| \geq c$ and equals zero otherwise. This establishes the following result.

Lemma 5.3 (Markov) *Let z be a random variable with finite $p^{\,th}$ moment. Then,*

$$\mathbb{P}(|z| \geq c) \leq \frac{\mathbb{E}\,|z|^p}{c^p},$$

where c is a positive real number.

For $p = 2$, Lemma 5.3 is also known as the *Chebyshev inequality*. If c is small such that $\mathbb{E}\,|z|^p / c^p > 1$, Markov's inequality is trivial. When c becomes large, the probability that z assumes very extreme values will be vanishing at the rate c^{-p}.

Another useful result in probability theory is stated below without proof.

Lemma 5.4 (Hölder) *Let y be a random variable with finite p^{th} moment ($p > 1$) and z a random variable with finite q^{th} moment ($q = p/(p-1)$). Then, $\mathbb{E}\,|yz| \le \|y\|_p \, \|z\|_q$.*

For $p = 2$, we have $\mathbb{E}\,|yz| \le \|y\|_2 \, \|z\|_2$. By noting that $|\,\mathbb{E}(yz)| < \mathbb{E}\,|yz|$, we immediately have the next result; cf. Lemma 2.3.

Lemma 5.5 (Cauchy-Schwartz) *Let y and z be two square integrable random variables. Then, $|\,\mathbb{E}(yz)| \le \|y\|_2 \, \|z\|_2$.*

Let $y = 1$ and $x = z^p$. Then for $q > p$ and $r = q/p$, Hölder's inequality also ensures that

$$\mathbb{E}\,|z^p| \le \|x\|_r \, \|y\|_{r/(r-1)} = [\mathbb{E}(z^{pr})]^{1/r} = [\mathbb{E}(z^q)]^{p/q}.$$

This shows that when a random variable has finite q^{th} moment, it must also have finite p^{th} moment for any $p < q$, as stated below.

Lemma 5.6 (Liapunov) *Let z be a random variable with finite q^{th} moment. Then for $p < q$, $\|z\|_p \le \|z\|_q$.*

The inequality below states that the L_p-norm of a finite sum is less than the sum of individual L_p-norms.

Lemma 5.7 (Minkowski) *Let z_i, $i = 1, \ldots, n$, be random variables with finite p^{th} moment ($p \ge 1$). Then, $\|\sum_{i=1}^n z_i\|_p \le \sum_{i=1}^n \|z_i\|_p$.*

When there are only two random variables in the sum, this is just the *triangle inequality* for L_p-norms; see also Exercise 5.3.

5.2 Conditional Distributions and Moments

Given two events A and B in $\mathcal{F}$, if it is known that B has occurred, the outcome space is restricted to B, so that the outcomes of A must be in $A \cap B$. The likelihood of A is thus characterized by the conditional probability

$$\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\,\mathbb{P}(B),$$

for $\mathbb{P}(B) \neq 0$. It can be shown that $\mathbb{P}(\cdot|B)$ satisfies the axioms for probability measures; see Exerise 5.4. This concept is readily extended to construct *conditional density function* and *conditional distribution function*.

5.2.1 Conditional Distributions

Let $\boldsymbol{y}$ and $\boldsymbol{z}$ denote two integrable random vectors such that $\boldsymbol{z}$ has the density function $f_{\boldsymbol{z}}$. For $f_{\boldsymbol{y}}(\boldsymbol{\eta}) \neq 0$, define the conditional density function of $\boldsymbol{z}$ given $\boldsymbol{y} = \boldsymbol{\eta}$ as

$$f_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta}) = \frac{f_{\boldsymbol{z},\boldsymbol{y}}(\boldsymbol{\zeta}, \boldsymbol{\eta})}{f_{\boldsymbol{y}}(\boldsymbol{\eta})},$$

which is clearly non-negative whenever it is defined. This function also integrates to one on $\mathbb{R}^d$ because

$$\int_{\mathbb{R}^d} f_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\zeta} = \frac{1}{f_{\boldsymbol{y}}(\boldsymbol{\eta})} \int_{\mathbb{R}^d} f_{\boldsymbol{z},\boldsymbol{y}}(\boldsymbol{\zeta}, \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\zeta} = \frac{1}{f_{\boldsymbol{y}}(\boldsymbol{\eta})} f_{\boldsymbol{y}}(\boldsymbol{\eta}) = 1.$$

Thus, $f_{\boldsymbol{z}|\boldsymbol{y}}$ is a legitimate density function. For example, the bivariate density function of two random variables z and y forms a surface on the zy-plane. By fixing $y = \eta$, we obtain a cross section (slice) under this surface. Dividing the joint density by the marginal density $f_y(\eta)$ amounts to adjusting the height of this slice so that the resulting area integrates to one.

Given the conditional density function $f_{\boldsymbol{z}|\boldsymbol{y}}$, we have for $A \in \mathcal{B}^d$,

$$\mathbb{P}(\boldsymbol{z} \in A \mid \boldsymbol{y} = \boldsymbol{\eta}) = \int_A f_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\zeta}.$$

Note that this conditional probability is defined even when $\mathbb{P}(\boldsymbol{y} = \boldsymbol{\eta})$ may be zero. In particular, when

$$A = (-\infty, \zeta_1] \times \cdots \times (-\infty, \zeta_d],$$

we obtain the *conditional distribution* function:

$$F_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta}) = \mathbb{P}(z_1 \leq \zeta_1, \ldots, z_d \leq \zeta_d \mid \boldsymbol{y} = \boldsymbol{\eta}).$$

When $\boldsymbol{z}$ and $\boldsymbol{y}$ are independent, the conditional density (distribution) simply reduces to the unconditional density (distribution).

5.2.2 Conditional Moments

Analogous to unconditional expectation, the *conditional expectation* of the integrable random variable z_i given the information $\boldsymbol{y} = \boldsymbol{\eta}$ is

$$\mathbb{E}(z_i \mid \boldsymbol{y} = \boldsymbol{\eta}) = \int_{\mathbb{R}} \zeta_i\, \mathrm{d}F_{\boldsymbol{z}|\boldsymbol{y}}(\zeta_i \mid \boldsymbol{y} = \boldsymbol{\eta});$$

the conditional expectation of the random vector $\boldsymbol{z}$ is $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y} = \boldsymbol{\eta})$ which is defined elementwise. By allowing $\boldsymbol{y}$ to vary across all possible values $\boldsymbol{\eta}$, we obtain the conditional

expectation function $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})$ whose value depends on $\boldsymbol{\eta}$, the realization of $\boldsymbol{y}$. Thus, $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})$ is a function of $\boldsymbol{y}$ and hence also a random vector.

More generally, the conditional expectation can be defined by taking a suitable σ-algebra as the conditioning set. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})$ is the integrable and $\mathcal{G}$-measurable random variable satisfying

$$\int_G \mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) \, \mathrm{d}\,\mathbb{P} = \int_G \boldsymbol{z} \, \mathrm{d}\,\mathbb{P}, \quad \forall G \in \mathcal{G}.$$

This definition basically says that the conditional expectation with respect to $\mathcal{G}$ is such that its weighted sum is the same as that of $\boldsymbol{z}$ over any G in $\mathcal{G}$. Suppose that $\mathcal{G}$ is the trivial σ-algebra $\{\Omega, \emptyset\}$, i.e., the smallest σ-algebra that contains no extra information from any random vectors. For the conditional expectation with respect to the trivial σ-algebra, it is readily seen that it must be a constant $\boldsymbol{c}$ with probability one so as to be measurable with respect to $\{\Omega, \emptyset\}$. Then,

$$\mathbb{E}(\boldsymbol{z}) = \int_\Omega \boldsymbol{z} \, \mathrm{d}\,\mathbb{P} = \int_\Omega \boldsymbol{c} \, \mathrm{d}\,\mathbb{P} = \boldsymbol{c}.$$

That is, the conditional expectation with respect to the trivial σ-algebra is the unconditional expectation $\mathbb{E}(\boldsymbol{z})$. Consider now $\mathcal{G} = \sigma(\boldsymbol{y})$, the σ-algebra generated by $\boldsymbol{y}$. We also write

$$\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}) = \mathbb{E}[\boldsymbol{z} \mid \sigma(\boldsymbol{y})],$$

which is interpreted as the prediction of $\boldsymbol{z}$ given all the information associated with $\boldsymbol{y}$.

Similar to unconditional expectations, conditional expectations are monotonic: if $z \geq x$ with probability one, then $\mathbb{E}(z \mid \mathcal{G}) \geq \mathbb{E}(x \mid \mathcal{G})$ with probability one; in particular, if z is non-negative with probability one, then $\mathbb{E}(z \mid \mathcal{G}) \geq 0$ with probability one. Moreover, If $\boldsymbol{z}$ is independent of $\boldsymbol{y}$, then $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}) = \mathbb{E}(\boldsymbol{z})$. For example, when $\boldsymbol{z}$ is a constant vector $\boldsymbol{c}$ which is independent of any random variable, $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}) = \boldsymbol{c}$. The linearity result below is analogous to Lemma 2.2 for unconditional expectations.

Lemma 5.8 *Let $\boldsymbol{z}$ $(d \times 1)$ and $\boldsymbol{y}$ $(c \times 1)$ be integrable random vectors and $\boldsymbol{A}$ $(n \times d)$ and $\boldsymbol{B}$ $(n \times c)$ be non-stochastic matrices. Then with probability one,*

$$\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{B}\boldsymbol{y} \mid \mathcal{G}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) + \boldsymbol{B}\,\mathbb{E}(\boldsymbol{y} \mid \mathcal{G}).$$

If $\boldsymbol{b}$ $(n \times 1)$ is a non-stochastic vector, $\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} \mid \mathcal{G}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) + \boldsymbol{b}$ with probability one.

From the definition of conditional expectation, we immediately have

$$\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})] = \int_\Omega \mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) \, \mathrm{d}\,\mathbb{P} = \int_\Omega \boldsymbol{z} \, \mathrm{d}\,\mathbb{P} = \mathbb{E}(\boldsymbol{z});$$

this is known as the *law of iterated expectations*. This result also suggests that if conditional expectations are taken sequentially with respect to a collection of nested σ-algebras, only the smallest σ-algebra matters. For example, for k random vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$,

$$\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}] = \mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}).$$

A formal result is stated below; see Exercise 5.5.

Lemma 5.9 (Law of Iterated Expectations) *Let $\mathcal{G}$ and $\mathcal{H}$ be two sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{G} \subseteq \mathcal{H}$. Then for the integrable random vector $\boldsymbol{z}$,*

$$\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{H}) \mid \mathcal{G}] = \mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) \mid \mathcal{H}] = \mathbb{E}(\boldsymbol{z} \mid \mathcal{G});$$

in particular, $\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})] = \mathbb{E}(\boldsymbol{z})$.

If $\boldsymbol{z}$ is $\mathcal{G}$-measurable, all the information resulted from $\boldsymbol{z}$ is already contained in $\mathcal{G}$ so that $\boldsymbol{z}$ can be treated as "known" in $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})$ and taken out from the conditional expectation. That is, $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) = \boldsymbol{z}$ with probability one. Hence,

$$\mathbb{E}(\boldsymbol{z}\boldsymbol{x}' \mid \mathcal{G}) = \boldsymbol{z}\,\mathbb{E}(\boldsymbol{x}' \mid \mathcal{G}).$$

In particular, $\boldsymbol{z}$ can be taken out from the conditional expectation when $\boldsymbol{z}$ itself is a conditioning variable. This result is generalized as follows.

Lemma 5.10 *Let $\boldsymbol{z}$ be a $\mathcal{G}$-measurable random vector. Then for any Borel-measurable function g,*

$$\mathbb{E}[g(\boldsymbol{z})\boldsymbol{x} \mid \mathcal{G}] = g(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{x} \mid \mathcal{G}),$$

with probability one.

Two square integrable random variables z and y are said to be *orthogonal* if their inner product $\mathbb{E}(zy) = 0$. This definition allows us to discuss orthogonal projection in the space of square integrable random vectors. Let z be a square integrable random variable and $\tilde{z}$ be a $\mathcal{G}$-measurable random variable. Then, by Lemma 5.9 (law of iterated expectations) and Lemma 5.10,

$$\begin{aligned}
\mathbb{E}\big[\big(z - \mathbb{E}(z \mid \mathcal{G})\big)\tilde{z}\big] &= \mathbb{E}\Big[\mathbb{E}\big[\big(z - \mathbb{E}(z \mid \mathcal{G})\big)\tilde{z} \mid \mathcal{G}\big]\Big] \\
&= \mathbb{E}\big[\mathbb{E}(z \mid \mathcal{G})\tilde{z} - \mathbb{E}(z \mid \mathcal{G})\tilde{z}\big] \\
&= 0.
\end{aligned}$$

That is, the difference between z and its conditional expectation $\mathbb{E}(z \mid \mathcal{G})$ must be orthogonal to any $\mathcal{G}$-measurable random variable. It can then be seen that for any square integrable, $\mathcal{G}$-measurable random variable $\tilde{z}$,

$$\begin{aligned}\mathbb{E}(z-\tilde{z})^2 &= \mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})+\mathbb{E}(z \mid \mathcal{G})-\tilde{z}]^2 \\ &= \mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})]^2+\mathbb{E}[\mathbb{E}(z \mid \mathcal{G})-\tilde{z}]^2 \\ &\geq \mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})]^2,\end{aligned}$$

where in the second equality the cross-product term vanishes because both $\mathbb{E}(z \mid \mathcal{G})$ and $\tilde{z}$ are $\mathcal{G}$-measurable and hence orthogonal to $z-\mathbb{E}(z \mid \mathcal{G})$. That is, among all $\mathcal{G}$-measurable random variables that are also square integrable, $\mathbb{E}(z \mid \mathcal{G})$ is the closest to z in terms of the L_2-norm. This shows that $\mathbb{E}(z \mid \mathcal{G})$ is the orthogonal projection of z onto the space of all $\mathcal{G}$-measurable, square integrable random variables.

Lemma 5.11 *Let z be a square integrable random variable. Then*

$$\mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})]^2 \leq \mathbb{E}(z-\tilde{z})^2,$$

for any $\mathcal{G}$-measurable random variable $\tilde{z}$.

In particular, let $\mathcal{G}=\sigma(\boldsymbol{y})$, where $\boldsymbol{y}$ is a square integrable random vector. Lemma 5.11 implies that

$$\mathbb{E}\big[z-\mathbb{E}\big(z \mid \sigma(\boldsymbol{y})\big)\big]^2 \leq \mathbb{E}\big(z-h(\boldsymbol{y})\big)^2,$$

for any Borel-measurable function h such that $h(\boldsymbol{y})$ is also square integrable. Thus, $\mathbb{E}[z \mid \sigma(\boldsymbol{y})]$ minimizes the L_2-norm $\|z-h(\boldsymbol{y})\|_2$, and its difference from z is orthogonal to any function of $\boldsymbol{y}$ that is also square integrable. We may then say that, given all the information generated from $\boldsymbol{y}$, $\mathbb{E}[z \mid \sigma(\boldsymbol{y})]$ is the "best approximation" of z in terms of the L_2-norm (or simply the best L_2 predictor).

The *conditional variance-covariance matrix* of $\boldsymbol{z}$ given $\boldsymbol{y}$ is

$$\begin{aligned}\operatorname{var}(\boldsymbol{z} \mid \boldsymbol{y}) &= \mathbb{E}\big([\boldsymbol{z}-\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})][\boldsymbol{z}-\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})]' \mid \boldsymbol{y}\big) \\ &= \mathbb{E}(\boldsymbol{z}\boldsymbol{z}' \mid \boldsymbol{y})-\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})\,\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})'.\end{aligned}$$

Similar to unconditional variance-covariance matrix, we have for non-stochastic matrices $\boldsymbol{A}$ and $\boldsymbol{b}$,

$$\operatorname{var}(\boldsymbol{A}\boldsymbol{z}+\boldsymbol{b} \mid \boldsymbol{y}) = \boldsymbol{A}\,\operatorname{var}(\boldsymbol{z} \mid \boldsymbol{y})\,\boldsymbol{A}',$$

which is nonsingular provided that $\boldsymbol{A}$ has full row rank and $\mathrm{var}(\boldsymbol{z} \mid \boldsymbol{y})$ is positive definite. It can also be shown that

$$\mathrm{var}(\boldsymbol{z}) = \mathbb{E}[\mathrm{var}(\boldsymbol{z} \mid \boldsymbol{y})] + \mathrm{var}\big(\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})\big);$$

see Exercise 5.6. That is, the variance of $\boldsymbol{z}$ can be expressed as the sum of two components: the mean of its conditional variance and the variance of its conditional mean. This is also known as the decomposition of *analysis of variance*.

Example 5.12 Suppose that $(\boldsymbol{y}' \; \boldsymbol{x}')'$ is distributed as a multivariate normal random vector:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_x} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_y} & \boldsymbol{\Sigma}'_{\boldsymbol{xy}} \\ \boldsymbol{\Sigma_{xy}} & \boldsymbol{\Sigma_x} \end{bmatrix} \right).$$

It is easy to see that the conditional density function of $\boldsymbol{y}$ given $\boldsymbol{x}$, obtained from dividing the multivariate normal density function of $\boldsymbol{y}$ and $\boldsymbol{x}$ by the normal density of $\boldsymbol{x}$, is also normal with the conditional mean

$$\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x}) = \boldsymbol{\mu_y} - \boldsymbol{\Sigma}'_{\boldsymbol{xy}} \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_x}),$$

and the conditional variance-covariance matrix

$$\mathrm{var}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathrm{var}(\boldsymbol{y}) - \mathrm{var}\big(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})\big) = \boldsymbol{\Sigma_y} - \boldsymbol{\Sigma}'_{\boldsymbol{xy}} \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma_{xy}}.$$

Note that $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$ is a linear function of $\boldsymbol{x}$ and that $\mathrm{var}(\boldsymbol{y} \mid \boldsymbol{x})$ does not vary with $\boldsymbol{x}$.

5.3 Modes of Convergence

Consider now a sequence of random variables $\{z_n(\omega)\}_{n=1,2,\ldots}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For a given ω, $\{z_n\}$ is a realization (a sequence of sample values) of the random element ω with the index n, and that for a given n, z_n is a random variable which assumes different values depending on ω. In this section we will discuss various modes of convergence for sequences of random variables.

5.3.1 Almost Sure Convergence

We first introduce the concept of *almost sure convergence* (*convergence with probability one*). Suppose that $\{z_n\}$ is a sequence of random variables and z is a random variable, all defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence $\{z_n\}$ is said to converge to z almost surely if, and only if,

$$\mathbb{P}(\omega \colon z_n(\omega) \to z(\omega) \text{ as } n \to \infty) = 1,$$

denoted as $z_n \xrightarrow{\text{a.s.}} z$ or $z_n \to z$ a.s. Note that for a given ω, the realization $z_n(\omega)$ may or may not converge to $z(\omega)$. Almost sure convergence requires that $z_n(\omega) \to z(\omega)$ for almost all $\omega \in \Omega$, except for those ω in a set with probability zero. That is, almost all the realizations $z_n(\omega)$ will be eventually close to $z(\omega)$ for all n sufficiently large; the event that z_n will not approach z is improbable. When $\boldsymbol{z}_n$ and $\boldsymbol{z}$ are both $\mathbb{R}^d$-valued, almost sure convergence is defined elementwise. That is, $\boldsymbol{z}_n \to \boldsymbol{z}$ a.s. if every element of $\boldsymbol{z}_n$ converges almost surely to the corresponding element of $\boldsymbol{z}$.

The following result shows that continuous transformation preserves almost sure convergence.

Lemma 5.13 *Let $g\colon \mathbb{R} \mapsto \mathbb{R}$ be a function continuous on $S_g \subseteq \mathbb{R}$.*

[a] *If $z_n \xrightarrow{\text{a.s.}} z$, where z is a random variable such that $\mathbb{P}(z \in S_g) = 1$, then $g(z_n) \xrightarrow{\text{a.s.}} g(z)$.*

[b] *If $z_n \xrightarrow{\text{a.s.}} c$, where c is a real number at which g is continuous, then $g(z_n) \xrightarrow{\text{a.s.}} g(c)$.*

Proof: Let $\Omega_0 = \{\omega\colon z_n(\omega) \to z(\omega)\}$ and $\Omega_1 = \{\omega\colon z(\omega) \in S_g\}$. Thus, for $\omega \in (\Omega_0 \cap \Omega_1)$, continuity of g ensures that $g(z_n(\omega)) \to g(z(\omega))$. Note that

$$(\Omega_0 \cap \Omega_1)^c = \Omega_0^c \cup \Omega_1^c,$$

which has probability zero because $\mathbb{P}(\Omega_0^c) = \mathbb{P}(\Omega_1^c) = 0$. It follows that $\Omega_0 \cap \Omega_1$ has probability one. This proves that $g(z_n) \to g(z)$ with probability one. The second assertion is just a special case of the first result. □

Lemma 5.13 is easily generalized to $\mathbb{R}^d$-valued random variables. For example, $\boldsymbol{z}_n \xrightarrow{\text{a.s.}} \boldsymbol{z}$ implies

$$\begin{aligned} z_{1,n} + z_{2,n} &\xrightarrow{\text{a.s.}} z_1 + z_2, \\ z_{1,n} z_{2,n} &\xrightarrow{\text{a.s.}} z_1 z_2, \\ z_{1,n}^2 + z_{2,n}^2 &\xrightarrow{\text{a.s.}} z_1^2 + z_2^2, \end{aligned}$$

where $z_{1,n}, z_{2,n}$ are two elements of $\boldsymbol{z}_n$ and z_1, z_2 are the corresponding elements of $\boldsymbol{z}$. Also, provided that $z_2 \neq 0$ with probability one, $z_{1,n}/z_{2,n} \to z_1/z_2$ a.s.

5.3.2 Convergence in Probability

A convergence concept that is weaker than almost sure convergence is *convergence in probability*. A sequence of random variables $\{z_n\}$ is said to converge to z in probability if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\omega\colon |z_n(\omega) - z(\omega)| > \epsilon) = 0,$$

or equivalently,

$$\lim_{n\to\infty} \mathbb{P}(\omega\colon |z_n(\omega) - z(\omega)| \le \epsilon) = 1,$$

denoted as $z_n \xrightarrow{\mathbb{P}} z$ or $z_n \to z$ in probability. We also say that z is the *probability limit* of z_n, denoted as $\operatorname{plim} z_n = z$. In particular, if the probability limit of z_n is a constant c, all the probability mass of z_n will concentrate around c when n becomes large. For $\mathbb{R}^d$-valued random variables $\boldsymbol{z}_n$ and $\boldsymbol{z}$, convergence in probability is also defined elementwise.

In the definition of convergence in probability, the events $\Omega_n(\epsilon) = \{\omega\colon |z_n(\omega) - z(\omega)| \le \epsilon\}$ vary with n, and convergence is referred to the probabilities of such events: $p_n = \mathbb{P}(\Omega_n(\epsilon))$, rather than the random variables z_n. By contrast, almost sure convergence is related directly to the behaviors of random variables. For convergence in probability, the event Ω_n that z_n will be close to z becomes highly likely when n tends to infinity, or its complement (z_n will deviate from z by a certain distance) becomes highly unlikely when n tends to infinity. Whether z_n will converge to z is not of any concern in convergence in probability.

More specifically, let Ω_0 denote the set of ω such that $z_n(\omega)$ converges to $z(\omega)$. For $\omega \in \Omega_0$, there is some m such that ω is in $\Omega_n(\epsilon)$ for all $n > m$. That is,

$$\Omega_0 \subseteq \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \Omega_n(\epsilon) \in \mathcal{F}.$$

As $\cap_{n=m}^{\infty}\Omega_n(\epsilon)$ is also in $\mathcal{F}$ and non-decreasing in m, it follows that

$$\mathbb{P}(\Omega_0) \le \mathbb{P}\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \Omega_n(\epsilon)\right) = \lim_{m\to\infty} \mathbb{P}\left(\bigcap_{n=m}^{\infty} \Omega_n(\epsilon)\right) \le \lim_{m\to\infty} \mathbb{P}\big(\Omega_m(\epsilon)\big).$$

This inequality proves that almost sure convergence implies convergence in probability, but the converse is not true in general. We state this result below.

Lemma 5.14 *If* $z_n \xrightarrow{\text{a.s.}} z$, *then* $z_n \xrightarrow{\mathbb{P}} z$.

The following well-known example shows that when there is convergence in probability, the random variables themselves may not even converge for any ω.

Example 5.15 Let $\Omega = [0,1]$ and $\mathbb{P}$ be the Lebesgue measure (i.e., $\mathbb{P}\{(a,b]\} = b - a$ for $(a,b] \subseteq [0,1]$). Consider the sequence $\{I_n\}$ of intervals $[0,1]$, $[0,1/2)$, $[1/2,1]$, $[0,1/3)$, $[1/3,2/3)$, $[2/3,1]$, $\ldots$, and let $z_n = \mathbf{1}_{I_n}$ be the indicator function of I_n: $z_n(\omega) = 1$ if $\omega \in I_n$ and $z_n = 0$ otherwise. When n tends to infinity, I_n shrinks toward a singleton which has the Lebesgue measure zero. For $0 < \epsilon < 1$, we then have

$$\mathbb{P}(|z_n| > \epsilon) = \mathbb{P}(I_n) \to 0,$$

which shows $z_n \xrightarrow{\mathbb{P}} 0$. On the other hand, it is easy to see that each $\omega \in [0,1]$ must be covered by infinitely many intervals. Thus, given any $\omega \in [0,1]$, $z_n(\omega) = 1$ for infinitely many n, and hence $z_n(\omega)$ does not converge to zero. Note that convergence in probability permits z_n to deviate from the probability limit infinitely often, but almost sure convergence does not, except for those ω in the set of probability zero. □

Intuitively, when z_n has finite variance such that $\mathrm{var}(z_n)$ vanishes asymptotically, the distribution of z_n would shrink toward its mean $\mathbb{E}(z_n)$. If, in addition, $\mathbb{E}(z_n)$ tends to a constant c (or $\mathbb{E}(z_n) = c$), then z_n ought to be degenerate at c in the limit. These observations suggest the following sufficient conditions for convergence in probability; see Exercises 5.7 and 5.8. In many cases, it is easier to establish convergence in probability by verifying these conditions.

Lemma 5.16 *Let $\{z_n\}$ be a sequence of square integrable random variables. If $\mathbb{E}(z_n) \to c$ and $\mathrm{var}(z_n) \to 0$, then $z_n \xrightarrow{\mathbb{P}} c$.*

Analogous to Lemma 5.13, continuous functions also preserve convergence in probability.

Lemma 5.17 *Let $g\colon \mathbb{R} \mapsto \mathbb{R}$ be a function continuous on $S_g \subseteq \mathbb{R}$.*

[a] *If $z_n \xrightarrow{\mathbb{P}} z$, where z is a random variable such that $\mathbb{P}(z \in S_g) = 1$, then $g(z_n) \xrightarrow{\mathbb{P}} g(z)$.*

[b] **(Slutsky)** *If $z_n \xrightarrow{\mathbb{P}} c$, where c is a real number at which g is continuous, then $g(z_n) \xrightarrow{\mathbb{P}} g(c)$.*

Proof: By the continuity of g, for each $\epsilon > 0$, we can find a $\delta > 0$ such that

$$\begin{aligned}\{\omega : |z_n(\omega) - z(\omega)| \le \delta\} \cap \{\omega : z(\omega) \in S_g\} \\ \subseteq \{\omega : |g(z_n(\omega)) - g(z(\omega))| \le \epsilon\}.\end{aligned}$$

Taking complementation of both sides and noting that the complement of $\{\omega : z(\omega) \in S_g\}$ has probability zero, we have

$$\mathbb{P}(|g(z_n) - g(z)| > \epsilon) \le \mathbb{P}(|z_n - z| > \delta).$$

As z_n converges to z in probability, the right-hand side converges to zero and so does the left-hand side. □

Lemma 5.17 is readily generalized to $\mathbb{R}^d$-valued random variables. For instance, $\boldsymbol{z}_n \xrightarrow{\mathbb{P}} \boldsymbol{z}$ implies

$$\begin{aligned} z_{1,n} + z_{2,n} &\xrightarrow{\mathbb{P}} z_1 + z_2, \\ z_{1,n} z_{2,n} &\xrightarrow{\mathbb{P}} z_1 z_2, \\ z_{1,n}^2 + z_{2,n}^2 &\xrightarrow{\mathbb{P}} z_1^2 + z_2^2, \end{aligned}$$

where $z_{1,n}, z_{2,n}$ are two elements of $\boldsymbol{z}_n$ and z_1, z_2 are the corresponding elements of $\boldsymbol{z}$. Also, provided that $z_2 \neq 0$ with probability one, $z_{1,n}/z_{2,n} \xrightarrow{\mathbb{P}} z_1/z_2$.

5.3.3 Convergence in Distribution

Another convergence mode, known as *convergence in distribution* or *convergence in law*, concerns the behavior of the distribution functions of random variables. Let F_{z_n} and F_z be the distribution functions of z_n and z, respectively. A sequence of random variables $\{z_n\}$ is said to converge to z in distribution, denoted as $z_n \xrightarrow{D} z$, if

$$\lim_{n\to\infty} F_{z_n}(\zeta) = F_z(\zeta),$$

for every continuity point ζ of F_z. That is, regardless the distributions of z_n, convergence in distribution ensures that F_{z_n} will be arbitrarily close to F_z for all n sufficiently large. The distribution F_z is thus known as the *limiting distribution* of z_n. We also say that z_n is asymptotically distributed as F_z, denoted as $z_n \overset{A}{\sim} F_z$.

For random vectors $\{\boldsymbol{z}_n\}$ and $\boldsymbol{z}$, $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ if the joint distributions $F_{\boldsymbol{z}_n}$ converge to $F_{\boldsymbol{z}}$ for every continuity point $\boldsymbol{\zeta}$ of $F_{\boldsymbol{z}}$. It is, however, more cumbersome to show convergence in distribution for a sequence of random vectors. The so-called *Cramér-Wold device* allows us to transform this multivariate convergence problem to a univariate one. This result is stated below without proof.

Lemma 5.18 (Cramér-Wold Device) *Let $\{\boldsymbol{z}_n\}$ be a sequence of random vectors in $\mathbb{R}^d$. Then $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ if and only if $\boldsymbol{\alpha}'\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{\alpha}'\boldsymbol{z}$ for every $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$.*

There is also a uni-directional relationship between convergence in probability and convergence in distribution. To see this, note that for some arbitrary $\epsilon > 0$ and a continuity point ζ of F_z, we have

$$\begin{aligned} \mathbb{P}(z_n \leq \zeta) &= \mathbb{P}(\{z_n \leq \zeta\} \cap \{|z_n - z| \leq \epsilon\}) + \mathbb{P}(\{z_n \leq \zeta\} \cap \{|z_n - z| > \epsilon\}) \\ &\leq \mathbb{P}(z \leq \zeta + \epsilon) + \mathbb{P}(|z_n - z| > \epsilon). \end{aligned}$$

Similarly,

$$\mathbb{P}(z \leq \zeta - \epsilon) \leq \mathbb{P}(z_n \leq \zeta) + \mathbb{P}(|z_n - z| > \epsilon).$$

If $z_n \xrightarrow{\mathbb{P}} z$, then by passing to the limit and noting that ϵ is arbitrary, the inequalities above imply

$$\lim_{n\to\infty} \mathbb{P}(z_n \leq \zeta) = \mathbb{P}(z \leq \zeta).$$

That is, $F_{z_n}(\zeta) \to F_z(\zeta)$. The converse is not true in general, however.

When z_n converges in distribution to a real number c, it is not difficult to show that z_n also converges to c in probability. In this case, these two convergence modes are equivalent. To be sure, note that a real number c can be viewed as a degenerate random variable with the distribution function:

$$F(\zeta) = \begin{cases} 0, & \zeta < c, \\ 1, & \zeta \geq c, \end{cases}$$

which is a step function with a jump point at c. When $z_n \xrightarrow{D} c$, all the probability mass of z_n will concentrate at c as n becomes large; this is precisely what $z_n \xrightarrow{\mathbb{P}} c$ means. More formally, for any $\epsilon > 0$,

$$\mathbb{P}(|z_n - c| > \epsilon) = 1 - [F_{z_n}(c+\epsilon) - F_{z_n}((c-\epsilon)^-)],$$

where $(c-\epsilon)^-$ denotes the point adjacent to and less than $c - \epsilon$. Now, $z_n \xrightarrow{D} c$ implies that $F_{z_n}(c+\epsilon) - F_{z_n}((c-\epsilon)^-)$ converges to one, so that $\mathbb{P}(|z_n - c| > \epsilon)$ converges to zero. We summarizes these results below.

Lemma 5.19 *If $z_n \xrightarrow{\mathbb{P}} z$, then$z_n \xrightarrow{D} z$. For a constant c, $z_n \xrightarrow{\mathbb{P}} c$ is equivalent to $z_n \xrightarrow{D} c$.*

The *continuous mapping theorem* below asserts that continuous functions preserve convergence in distribution; cf. Lemmas 5.13 and 5.17.

Lemma 5.20 (Continuous Mapping Theorem) *Let $g\colon \mathbb{R} \mapsto \mathbb{R}$ be a function continuous almost everywhere on $\mathbb{R}$, except for at most countably many points. If $z_n \xrightarrow{D} z$, then $g(z_n) \xrightarrow{D} g(z)$.*

For example, if z_n converges in distribution to the standard normal random variable, the limiting distribution of z_n^2 is $\chi^2(1)$. Generalizing this result to $\mathbb{R}^d$-valued random variables,

we can see that when $\boldsymbol{z}_n$ converges in distribution to the d-dimensional standard normal random variable, the limiting distribution of $\boldsymbol{z}_n'\boldsymbol{z}_n$ is $\chi^2(d)$.

Two sequences of random variables $\{y_n\}$ and $\{z_n\}$ are said to be *asymptotically equivalent* if their differences $y_n - z_n$ converge to zero in probability. Intuitively, the limiting distributions of two asymptotically equivalent sequences, if exist, ought to be the same. This is stated in the next result without proof.

Lemma 5.21 *Let $\{y_n\}$ and $\{z_n\}$ be two sequences of random vectors such that $y_n - z_n \xrightarrow{\mathbb{P}} 0$. If $z_n \xrightarrow{D} z$, then $y_n \xrightarrow{D} z$.*

The next result is concerned with two sequences of random variables such that one converges in distribution and the other converges in probability.

Lemma 5.22 *If y_n converges in probability to a constant c and z_n converges in distribution to z, then $y_n + z_n \xrightarrow{D} c + z$, $y_n z_n \xrightarrow{D} cz$, and $z_n/y_n \xrightarrow{D} z/c$ if $c \neq 0$.*

5.4 Stochastic Order Notations

It is typical to use *order notations* to describe the behavior of a sequence of numbers, whether it converges or not. Let $\{c_n\}$ denote a sequence of positive real numbers.

1. Given a sequence $\{b_n\}$, we say that b_n is (at most) of order c_n, denoted as $b_n = O(c_n)$, if there exists a $\Delta < \infty$ such that $|b_n|/c_n \leq \Delta$ for all sufficiently large n. When c_n diverges, b_n cannot diverge faster than c_n; when c_n converges to zero, the rate of convergence of b_n is no slower than that of c_n. For example, the polynomial $a + bn$ is $O(n)$, and the partial sum of a bounded sequence $\sum_{i=1}^n b_i$ is $O(n)$. Note that an $O(1)$ sequence is a bounded sequence.

2. Given a sequence $\{b_n\}$, we say that b_n is of smaller order than c_n, denoted as $b_n = o(c_n)$, if $b_n/c_n \to 0$. When c_n diverges, b_n must diverge slower than c_n; when c_n converges to zero, the rate of convergence of b_n should be faster than that of c_n. For example, the polynomial $a+bn$ is $o(n^{1+\delta})$ for any $\delta > 0$, and the partial sum $\sum_{i=1}^n \alpha^i$, $|\alpha| < 1$, is $o(n)$. Note that an $o(1)$ sequence is a sequence that converges to zero.

If $\boldsymbol{b}_n$ is a vector (matrix), $\boldsymbol{b}_n$ is said to be $O(c_n)$ ($o(c_n)$) if every element of $\boldsymbol{b}_n$ is $O(c_n)$ ($o(c_n)$). It is also easy to verify the following results; see Exercise 5.10.

Lemma 5.23 *Let $\{a_n\}$ and $\{b_n\}$ be two non-stochastic sequences.*

(a) *If $a_n = O(n^r)$ and $b_n = O(n^s)$, then $a_n b_n = O(n^{r+s})$ and $a_n + b_n = O(n^{\max(r,s)})$.*

(b) *If* $a_n = o(n^r)$ *and* $b_n = o(n^s)$, *then* $a_n b_n = o(n^{r+s})$ *and* $a_n + b_n = o(n^{\max(r,s)})$.

(c) *If* $a_n = O(n^r)$ *and* $b_n = o(n^s)$, *then* $a_n b_n = o(n^{r+s})$ *and* $a_n + b_n = O(n^{\max(r,s)})$.

The order notations can be easily extended to describe the behavior of sequences of random variables. A sequence of random variables $\{z_n\}$ is said to be $O_{\text{a.s.}}(c_n)$ (or $O(c_n)$ almost surely) if z_n/c_n is $O(1)$ a.s., and it is said to be $O_{\mathbb{P}}(c_n)$ (or $O(c_n)$ in probability) if for every $\epsilon > 0$, there is some Δ such that

$$\mathbb{P}(|z_n|/c_n \geq \Delta) \leq \epsilon,$$

for all n sufficiently large. Similarly, $\{z_n\}$ is $o_{\text{a.s.}}(c_n)$ (or $o(c_n)$ almost surely) if $z_n/c_n \xrightarrow{\text{a.s.}} 0$, and it is $o_{\mathbb{P}}(c_n)$ (or $o(c_n)$ in probability) if $z_n/c_n \xrightarrow{\mathbb{P}} 0$.

If $\{z_n\}$ is $O_{\text{a.s.}}(1)$ ($o_{\text{a.s}}(1)$), we say that z_n is bounded (vanishing) almost surely; if $\{z_n\}$ is $O_{\mathbb{P}}(1)$ ($o_{\mathbb{P}}(1)$), z_n is bounded (vanishing) in probability. Note that Lemma 5.23 also holds for stochastic order notations. In particular, if a sequence of random variables is bounded almost surely (in probability) and another sequence of random variables is vanishing almost surely (in probability), the products of their corresponding elements are vanishing almost surely (in probability). That is, $y_n = O_{\text{a.s.}}(1)$ and $z_n = o_{\text{a.s.}}(1)$, then $y_n z_n$ is $o_{\text{a.s.}}(1)$.

When $z_n \xrightarrow{D} z$, we know that z_n does not converge in probability to z in general, but more can be said about the behavior of z_n. Let ζ be a continuity point of F_z. Then for any $\epsilon > 0$, we can choose a sufficiently large ζ such that $\mathbb{P}(|z| > \zeta) < \epsilon/2$. As $z_n \xrightarrow{D} z$, we can also choose n large enough such that

$$\mathbb{P}(|z_n| > \zeta) - \mathbb{P}(|z| > \zeta) < \epsilon/2,$$

which implies $\mathbb{P}(|z_n| > \zeta) < \epsilon$. This leads to the following conclusion.

Lemma 5.24 *Let* $\{z_n\}$ *be a sequence of random vectors such that* $z_n \xrightarrow{D} z$. *Then* $z_n = O_{\mathbb{P}}(1)$.

5.5 Law of Large Numbers

We first discuss the *law of large numbers* which is concerned with the averaging behavior of random variables. Intuitively, a sequence of random variables obeys a law of large numbers when its sample average essentially follows its mean behavior; random irregularities (deviations from the mean) are "wiped out" in the limit by averaging. When a law of large numbers holds almost surely, it is a *strong law of large numbers* (SLLN); when a law of large

numbers holds in probability, it is a *weak law of large numbers* (WLLN). For a sequence of random vectors (matrices), a SLLN (WLLN) is defined elementwise.

There are different versions of the SLLN (WLLN) for various types of random variables. Below is a well known SLLN for i.i.d. random variables.

Lemma 5.25 (Kolmogorov) *Let $\{z_t\}$ be a sequence of i.i.d. random variables with mean μ_o. Then,*

$$\frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\text{a.s.}} \mu_o.$$

This result asserts that, when z_t have a finite, common mean μ_o, the sample average of z_t is essentially close to μ_o, a non-stochastic number. Note, however, that i.i.d. random variables need not obey Kolmogorov's SLLN if they do not have a finite mean; for instance, Lemma 5.25 does not apply to i.i.d. Cauchy random variables. As almost sure convergence implies convergence in probability, the same condition in Lemma 5.25 ensures that $\{z_t\}$ also obeys a WLLN.

When $\{z_t\}$ is a sequence of independent random variables with possibly heterogeneous distributions, it may still obey a SLLN (WLLN) under a stronger condition.

Lemma 5.26 (Markov) *Let $\{z_t\}$ be a sequence of independent random variables with non-degenerate distributions such that for some $\delta > 0$, $\mathbb{E}\,|z_t|^{1+\delta}$ is bounded for all t. Then,*

$$\frac{1}{T}\sum_{t=1}^{T}[z_t - \mathbb{E}(z_t)] \xrightarrow{\text{a.s.}} 0,$$

Comparing to Kolmogorov's SLLN, Lemma 5.26 requires a stronger moment condition: bounded $(1+\delta)^{\text{th}}$ moment, yet z_t need not have a common mean. This SLLN indicates that the sample average of z_t eventually behaves like the average of $\mathbb{E}(z_t)$. Note that the average of $\mathbb{E}(z_t)$ may or may not converge; if it does converge to, say, μ^*,

$$\frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\text{a.s.}} \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(z_t) =: \mu^*.$$

Finally, as non-stochastic numbers can be viewed as independent random variables with degenerate distributions, it is understood that a non-stochastic sequence obeys a SLLN if its sample average converges.

The following example shows that a sequence of correlated random variables may also obey a WLLN.

Example 5.27 Suppose that y_t is generated as a weakly stationary AR(1) process:

$$y_t = \alpha_o y_{t-1} + u_t, \qquad |\alpha_o| < 1,$$

where u_t are i.i.d. random variables with mean zero and variance σ_u^2. In view of Section 4.3, we have $\mathbb{E}(y_t) = 0$, $\mathrm{var}(y_t) = \sigma_u^2/(1-\alpha_o^2)$, and

$$\mathrm{cov}(y_t, y_{t-j}) = \alpha_o^j \frac{\sigma_u^2}{1-\alpha_o^2}.$$

These results imply that $\mathbb{E}(T^{-1}\sum_{t=1}^T y_t) = 0$ and

$$\begin{aligned}
\mathrm{var}\left(\sum_{t=1}^T y_t\right) &= \sum_{t=1}^T \mathrm{var}(y_t) + 2\sum_{\tau=1}^{T-1}(T-\tau)\,\mathrm{cov}(y_t, y_{t-\tau}) \\
&\le \sum_{t=1}^T \mathrm{var}(y_t) + 2T\sum_{\tau=1}^{T-1}|\,\mathrm{cov}(y_t, y_{t-\tau})| \\
&= O(T).
\end{aligned}$$

The latter result shows that $\mathrm{var}\left(T^{-1}\sum_{t=1}^T y_t\right) = O(T^{-1})$ which converges to zero as T approaches infinity. It follows from Lemma 5.16 that

$$\frac{1}{T}\sum_{t=1}^T y_t \xrightarrow{\mathbb{P}} 0;$$

that is, $\{y_t\}$ obeys a WLLN. It can be seen that a key condition in the proof above is that the variance of $\sum_{t=1}^T y_t$ does not grow too rapidly (it is $O(T)$). The facts that y_t has a constant variance and that $\mathrm{cov}(y_t, y_{t-j})$ goes to zero exponentially fast as j tends to infinity are sufficient for this condition. This WLLN result is readily generalized to weakly stationary AR(p) processes. □

The example above shows that it may be quite cumbersome to establish a WLLN for weakly stationary processes. The lemma below gives a strong law for correlated random variables and is convenient in practice; see Davidson (1994, p. 326) for a more general result.

Lemma 5.28 *Let $y_t = \sum_{i=0}^\infty \pi_i u_{t-i}$, where u_t are i.i.d. random variables with mean zero and variance σ_u^2. If $\sum_{i=-\infty}^\infty |\pi_i| < \infty$, then $T^{-1}\sum_{t=1}^T y_t \xrightarrow{\text{a.s.}} 0$.*

In Example 5.27, $y_t = \sum_{i=0}^\infty \alpha_o^i u_{t-i}$ with $|\alpha_o| < 1$, so that $\sum_{i=0}^\infty |\alpha_o^i| < \infty$. Hence, Lemma 5.28 ensures that the average of y_t also converges to its mean (zero) almost surely. If $y_t = z_t - \mu_o$, then the average of z_t converges to $\mathbb{E}(z_t) = \mu_o$ almost surely. Comparing

to Example 5.27, Lemma 5.28 is quite general and applicable to any process that can be expressed as an MA process with absolutely summable weights.

From Lemmas 5.25, 5.26 and 5.28 we can see that a SLLN (WLLN) does not always hold. The random variables in a sequence must be "well behaved" (i.e., satisfying certain regularity conditions) to ensure a SLLN (WLLN). In particular, the sufficient conditions for a SLLN (WLLN) usually regulate the moments and dependence structure of random variables. Intuitively, random variables without certain bounded moment may exhibit aberrant behavior so that their random irregularities cannot be completely averaged out. For random variables with strong correlations over time, the variation of their partial sums may grow too rapidly and cannot be eliminated by simple averaging. More generally, it is also possible for a sequence of weakly dependent and heterogeneously distributed random variables to obey a SLLN (WLLN). This usually requires even stronger conditions on their moments and dependence structure. To avoid technicality, we will not give a SLLN (WLLN) for such general sequences but refer to White (2001) and Davidson (1994) for details. The following examples illustrate why a SLLN (WLLN) may fail to hold.

Example 5.29 Consider the sequences $\{t\}$ and $\{t^2\}$, $t = 1, 2, \ldots$. It is well known that

$$\sum_{t=1}^{T} t = T(T+1)/2, \qquad \sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6.$$

Hence, $T^{-1}\sum_{t=1}^{T} t$ and $T^{-1}\sum_{t=1}^{T} t^2$ both diverge. □

Example 5.30 Suppose that u_t are i.i.d. random variables with mean zero and variance σ_u^2. Thus, $T^{-1}\sum_{t=1}^{T} u_t \xrightarrow{\text{a.s.}} 0$ by Kolmogorv's SLLN (Lemma 5.25). Consider now $\{tu_t\}$. This sequence does not have bounded $(1+\delta)^{\text{th}}$ moment because $\mathbb{E}\,|tu_t|^{1+\delta}$ grows with t and therefore does not obey Markov's SLLN (Lemma 5.26). Moreover, note that

$$\operatorname{var}\left(\sum_{t=1}^{T} tu_t\right) = \sum_{t=1}^{T} t^2 \operatorname{var}(u_t) = \sigma_u^2 \frac{T(T+1)(2T+1)}{6}.$$

By Exercise 5.11, $\sum_{t=1}^{T} tu_t = O_{\mathbb{P}}(T^{3/2})$. It follows that $T^{-1}\sum_{t=1}^{T} tu_t = O_{\mathbb{P}}(T^{1/2})$, which shows that $\{tu_t\}$ does not obey a WLLN. □

Example 5.31 Suppose that y_t is generated as a *random walk*:

$$y_t = y_{t-1} + u_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, where u_t are i.i.d. random variables with mean zero and variance σ_u^2. Clearly,

$$y_t = \sum_{i=1}^{t} u_i,$$

which has mean zero and unbounded variance $t\sigma_u^2$. For $s < t$, write

$$y_t = y_s + \sum_{i=s+1}^{t} u_i = y_s + v_{t-s},$$

where $v_{t-s} = \sum_{i=s+1}^{t} u_i$ is independent of y_s. We then have

$$\operatorname{cov}(y_t, y_s) = \mathbb{E}(y_s^2) = s\sigma_u^2,$$

for $t > s$. Consequently,

$$\operatorname{var}\left(\sum_{t=1}^{T} y_t\right) = \sum_{t=1}^{T} \operatorname{var}(y_t) + 2\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} \operatorname{cov}(y_t, y_{t-\tau}).$$

It can be verified that the first term on the right-hand side is

$$\sum_{t=1}^{T} \operatorname{var}(y_t) = \sum_{t=1}^{T} t\sigma_u^2 = O(T^2),$$

and that the second term is

$$2\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} \operatorname{cov}(y_t, y_{t-\tau}) = 2\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} (t-\tau)\sigma_u^2 = O(T^3).$$

Thus, $\operatorname{var}(\sum_{t=1}^{T} y_t) = O(T^3)$, so that $\sum_{t=1}^{T} y_t = O_{\mathbb{P}}(T^{3/2})$ by Exercise 5.11. This shows that

$$\frac{1}{T}\sum_{t=1}^{T} y_t = O_{\mathbb{P}}(T^{1/2}),$$

which diverges in probability. This shows that when $\{y_t\}$ is a random walk, it does not obey a WLLN. In this case, y_t have unbounded variances and strong correlations over time. Due to these correlations, the variation of the partial sum of y_t grows much too fast. (Recall that the variance of $\sum_{t=1}^{T} y_t$ is only $O(T)$ in Example 5.27.) The conclusions above will not be altered when $\{u_t\}$ is a white noise or a weakly stationary process. □

Example 5.32 Suppose that y_t is generated as a random walk:

$$y_t = y_{t-1} + u_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, as in Example 5.31. Then, the sequence $\{y_{t-1}u_t\}$ has mean zero and

$$\operatorname{var}(y_{t-1}u_t) = \mathbb{E}(y_{t-1}^2)\ \mathbb{E}(u_t^2) = (t-1)\sigma_u^4.$$

More interestingly, it can be seen that for $s < t$,

$$\operatorname{cov}(y_{t-1}u_t, y_{s-1}u_s) = \mathbb{E}(y_{t-1}y_{s-1}u_s)\ \mathbb{E}(u_t) = 0.$$

We then have

$$\operatorname{var}\left(\sum_{t=1}^{T} y_{t-1}u_t\right) = \sum_{t=1}^{T} \operatorname{var}(y_{t-1}u_t) = \sum_{t=1}^{T}(t-1)\sigma_u^4 = O(T^2),$$

and $\sum_{t=1}^{T} y_{t-1}u_t = O_{\mathbb{P}}(T)$. Note, however, that $\operatorname{var}(T^{-1}\sum_{t=1}^{T} y_{t-1}u_t)$ converges to $\sigma_u^4/2$, rather than 0. Thus, $T^{-1}\sum_{t=1}^{T} y_{t-1}u_t$ cannot behave like a non-stochastic number in the limit. This shows that $\{y_{t-1}u_t\}$ does not obey a WLLN, even though its partial sums are $O_{\mathbb{P}}(T)$. □

In the asymptotic analysis of ecnometric estimators and test statistics, we usually encounter functions of several random variables, e.g., the product of two random variables. In some cases, it is easy to find sufficient conditions ensuring a SLLN (WLLN) for these functions. For example, suppose that $z_t = x_t y_t$, where $\{x_t\}$ and $\{y_t\}$ are two mutually independent sequences of independent random variables, each with bounded $(2+\delta)^{\text{th}}$ moment. Then, z_t are also independent random variables and have bounded $(1+\delta)^{\text{th}}$ moment by the Cauchy-Schwartz inequality. Lemma 5.26 then provides the SLLN for $\{z_t\}$. When $\{x_t\}$ and $\{y_t\}$ are two sequences of correlated (or weakly dependent) random variables, it is more cumbersome to find suitable conditions on x_t and y_t that ensure a SLLN (WLLN).

In what follows, a sequence of integrable random variables z_t is said to obey a SLLN if

$$\frac{1}{T}\sum_{t=1}^{T}[z_t - \mathbb{E}(z_t)] \xrightarrow{\text{a.s.}} 0; \tag{5.1}$$

it is said to obey a WLLN if the almost sure convergence above is replaced by convergence in probability. When $\mathbb{E}(z_t)$ is a constant μ_o, (5.1) simplifies to

$$\frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\text{a.s.}} \mu_o.$$

In our analysis, we may only invoke this generic SLLN (WLLN).

5.6 Uniform Law of Large Numbers

It is also common to deal with functions of random variables and model parameters. For example, $q(z_t(\omega);\theta)$ is a random variable for a given parameter θ, and it is a function of θ for a given ω. When θ is fixed, we may impose conditions on q and z_t such that $\{q(z_t(\omega);\theta)\}$ obeys a SLLN (WLLN), as discussed in Section 5.5. When θ assumes values in the parameter space Θ, a SLLN (WLLN) that does not depend on θ is then needed.

More specifically, suppose that $\{q(z_t;\theta)\}$ obeys a SLLN for *each* $\theta \in \Theta$:

$$Q_T(\omega;\theta) = \frac{1}{T}\sum_{t=1}^{T} q(z_t(\omega);\theta) \xrightarrow{\text{a.s.}} Q(\theta),$$

where $Q(\theta)$ is a non-stochastic function of θ. As this convergent behavior may depend on θ, $\Omega_0^c(\theta) = \{\omega\colon Q_T(\omega;\theta) \not\to Q(\theta)\}$ varies with θ. When Θ is an interval of $\mathbb{R}$, $\cup_{\theta\in\Theta}\Omega_0^c(\theta)$ is an uncountable union of non-convergence sets and hence may not have probability zero, even though each $\Omega_0^c(\theta)$ does. Thus, the event that $Q_T(\omega;\theta) \to Q(\theta)$ for *all* θ, i.e., $\cap_{\theta\in\Theta}\Omega_0(\theta)$, may occur with probability less than one. In fact, the union of all $\Omega_0^c(\theta)$ may not even be in $\mathcal{F}$ (only countable unions of the elements in $\mathcal{F}$ are guaranteed to be in $\mathcal{F}$). If so, we cannot conclude anything regarding the convergence of $Q_T(\omega;\theta)$. Worse still is when θ also depends on T, as in the case where θ is replaced by an estimator $\tilde{\theta}_T$. There may not exist a finite T^* such that $Q_T(\omega;\tilde{\theta}_T)$ are arbitrarily close to $Q(\omega;\tilde{\theta}_T)$ for all $T > T^*$.

These observations suggest that we should study convergence that is *uniform* on the parameter space Θ. In particular, $Q_T(\omega;\theta)$ converges to $Q(\theta)$ uniformly in θ almost surely (in probability) if the largest possible difference:

$$\sup_{\theta\in\Theta} |Q_T(\theta) - Q(\theta)| \to 0, \quad \text{a.s. (in probability)}.$$

In what follows we always assume that this supremum is a random variables for all T. The example below, similar to Example 2.14 of Davidson (1994), illustrates the difference between uniform and pointwise convergence.

Example 5.33 Let z_t be i.i.d. random variables with zero mean and

$$q_T(z_t(\omega);\theta) = z_t(\omega) + \begin{cases} T\theta, & 0 \le \theta \le \frac{1}{2T}, \\ 1 - T\theta, & \frac{1}{2T} < \theta \le \frac{1}{T}, \\ 0, & \frac{1}{T} < \theta < \infty. \end{cases}$$

Observe that for $\theta \ge 1/T$ and $\theta = 0$,

$$Q_T(\omega;\theta) = \frac{1}{T}\sum_{t=1}^{T} q_T(z_t;\theta) = \frac{1}{T}\sum_{t=1}^{T} z_t,$$

which converges to zero almost surely by Kolmogorov's SLLN. Thus, for a given θ, we can always choose T large enough such that $Q_T(\omega;\theta) \xrightarrow{\text{a.s.}} 0$, where 0 is the pointwise limit. On the other hand, it can be seen that $\Theta = [0,\infty)$ and

$$\sup_{\theta\in\Theta} |Q_T(\omega;\theta)| = |\bar{z}_T + 1/2| \xrightarrow{\text{a.s.}} 1/2,$$

so that the uniform limit is different from the pointwise limit. □

Let z_{Tt} denote the t^{th} random variable in a sample of T variables. These random variables are indexed by both T and t and form a *triangular array*. In this array, there is only one random variable z_{11} when $T = 1$, there are two random variables z_{21} and z_{22} when $T = 2$, there are three random variables z_{31}, z_{32} and z_{33} when $T = 3$, and so on. If this array does not depend on T, it is simply a sequence of random variables. We now consider a triangular array of functions $q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})$, $t = 1, 2, \ldots, T$, where $\boldsymbol{z}_t$ are integrable random vectors and $\boldsymbol{\theta}$ is the parameter vector taking values in the parameter space $\Theta \in \mathbb{R}^m$. For notation simplicity, we will not explicitly write ω in the functions. We say that $\{q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})\}$ obeys a *strong uniform law of large numbers* (SULLN) if

$$\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T}[q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}) - \mathbb{E}(q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}))] \xrightarrow{\text{a.s.}} 0, \tag{5.2}$$

cf. (5.1). Similarly, $\{q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})\}$ is said to obey a *weak uniform law of large numbers* (WULLN) if the convergence condition above holds in probability. If q_{Tt} is $\mathbb{R}^m$-valued functions, the SULLN (WULLN) is defined elementwise.

We have seen that pointwise convergence alone does not imply uniform convergence. An interesting question one would ask is: What are the additional conditions required to guarantee uniform convergence? Let

$$Q_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}[q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}) - \mathbb{E}(q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}))].$$

Suppose that Q_T satisfies the following Lipschitz-type continuity requirement: for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\dagger$ in Θ,

$$|Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \le C_T\|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| \quad \text{a.s.},$$

where $\|\cdot\|$ denotes the Euclidean norm, and C_T is a random variable bounded almost surely and does not depend on $\boldsymbol{\theta}$. Under this condition, $Q_T(\boldsymbol{\theta}^\dagger)$ can be made arbitrarily close to $Q_T(\boldsymbol{\theta})$, provided that $\boldsymbol{\theta}^\dagger$ is sufficiently close to $\boldsymbol{\theta}$. Using the triangle inequality and taking supremum over $\boldsymbol{\theta}$ we have

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta})| \le \sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| + |Q_T(\boldsymbol{\theta}^\dagger)|.$$

Let Δ denote an almost sure bound of C_T. Then given any $\epsilon > 0$, choosing $\boldsymbol{\theta}^\dagger$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| < \epsilon/(2\Delta)$ implies

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \le C_T\,\frac{\epsilon}{2\Delta} \le \frac{\epsilon}{2},$$

uniformly in T. Moreover, because $Q_T(\boldsymbol{\theta})$ converges to 0 almost surely for each $\boldsymbol{\theta}$ in Θ, $|Q_T(\boldsymbol{\theta}^\dagger)|$ is also less than $\epsilon/2$ for sufficiently large T. Consequently,

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta})| \leq \epsilon,$$

for all T sufficiently large. As these results hold almost surely, we have a SULLN for $Q_T(\boldsymbol{\theta})$; the conditions ensuring a WULLN are analogous.

Lemma 5.34 *Suppose that for each $\boldsymbol{\theta} \in \Theta$, $\{q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})\}$ obeys a SLLN (WLLN) and that for $\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger \in \Theta$,*

$$|Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \leq C_T\|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| \quad a.s.,$$

where C_T is a random variable bounded almost surely (in probability) and does not depend on $\boldsymbol{\theta}$. Then, $\{q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})\}$ obeys a SULLN (WULLN).

Lemma 5.34 is quite convenient for establishing a SULLN (WULLN) because it requires only two conditions. First, the random functions must obey a standard SLLN (WLLN) for each $\boldsymbol{\theta}$ in the parameter space. Second, the function q_{Tt} must satisfy a Lipschitz-type continuity condition which amounts to requiring q_{Tt} to be sufficiently "smooth" in the second argument. Note, however, that C_T being bounded almost surely may imply that the random variables in q_{Tt} are also bounded almost surely. This requirement is much too restrictive in applications. Hence, a SULLN may not be readily obtained from Lemma 5.34. On the other hand, a WULLN is practically more plausible because the requirement that C_T is $O_{\mathbb{P}}(1)$ is much weaker. For example, the boundedness of $\mathbb{E}\,|C_T|$ is sufficient for C_T being $O_{\mathbb{P}}(1)$ by Markov's inequality. For more specific conditions ensuring these requirements we refer to Gallant and White (1988) and Bierens (1994).

5.7 Central Limit Theorem

The *central limit theorem* (CLT) is another important result in probability theory. When a CLT holds, the distributions of suitably normalized averages of random variables are close to the standard normal distribution in the limit, regardless of the original distributions of these random variables. This is a very powerful result in applications because, as far as the approximation of normalized sample averages is concerned, only the standard normal distribution matters.

There are also different versions of CLT for various types of random variables. The following CLT applies to i.i.d. random variables.

Lemma 5.35 (Lindeberg-Lévy) *Let $\{z_t\}$ be a sequence of i.i.d. random variables with mean μ_o and variance $\sigma_o^2 > 0$. Then,*

$$\frac{\sqrt{T}(\bar{z}_T - \mu_o)}{\sigma_o} \xrightarrow{D} \mathcal{N}(0,\, 1).$$

A sequence of i.i.d. random variables need not obey this CLT if they do not have a finite variance, e.g., random variables with $t(2)$ distribution. Comparing to Lemma 5.25, one can immediately see that the Lindeberg-Lévy CLT requires a stronger condition (i.e., finite variance) than does Kolmogorov's SLLN.

Remark: In this example, $\bar{z}_T$ converges to μ_o in probability, and its variance σ_o^2/T vanishes when T tends to infinity. To prevent a degenerate distribution in the limit, it is natural to consider the normalized average $T^{1/2}(\bar{z}_T - \mu_o)$, which has a constant variance σ_o^2 for all T. This explains why the normalizing factor $T^{1/2}$ is needed. For a normalizing factor T^a with $a < 1/2$, the normalized average still converges to zero because its variance vanishes in the limit. For a normalizing factor T^a with $a > 1/2$, the normalized average diverges. In both cases, the resulting normalized averages cannot have a well-behaved, non-degenerate distribution in the limit. Thus, when $\{z_t\}$ obeys a CLT, $\bar{z}_T$ is said to converge to μ_o at the rate $T^{-1/2}$.

Independent random variables may also have the effect of a CLT. Below is a a version of Liapunov's CLT for independent (but not necessarily identically distributed) random variables.

Lemma 5.36 *Let $\{z_{Tt}\}$ be a triangular array of independent random variables with mean μ_{Tt} and variance $\sigma_{Tt}^2 > 0$ such that*

$$\bar{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T} \sigma_{Tt}^2 \to \sigma_o^2 > 0.$$

If for some $\delta > 0$, $\mathbb{E}\,|z_{Tt}|^{2+\delta}$ are bounded for all t, then

$$\frac{\sqrt{T}(\bar{z}_T - \bar{\mu}_T)}{\sigma_o} \xrightarrow{D} \mathcal{N}(0, 1).$$

Note that this result requires a stronger condition (bounded $(2+\delta)^{\text{th}}$ moment) than does Markov's SLLN, Lemma 5.26. Comparing to Lindeberg-Lévy's CLT, Lemma 5.36 allows mean and variance to vary with t at the expense of a stronger moment condition.

The sufficient conditions for a CLT are similar to but usually stronger than those for a WLLN. In particular, the random variables that obey a CLT have bounded moment up

to some higher order and are asymptotically independent with dependence vanishing sufficiently fast. Moreover, every random variable must also be asymptotically negligible, in the sense that no random variable is influential in affecting the partial sums. Although we will not specify the regularity conditions explicitly, we note that weakly stationary AR and MA processes obey a CLT in general. A sequence of weakly dependent and heterogeneously distributed random variables may also obey a CLT, depending on its moment and dependence structure. The following examples show that a CLT may not always hold.

Example 5.37 Suppose that $\{u_t\}$ is a sequence of independent random variables with mean zero, variance σ_u^2, and bounded $(2+\delta)^{\text{th}}$ moment. From Example 5.29, we know $\text{var}(\sum_{t=1}^T tu_t)$ is $O(T^3)$, which implies that variance of $T^{-1/2}\sum_{t=1}^T tu_t$ is diverging at the rate $O(T^2)$. On the other hand, observe that

$$\text{var}\left(\frac{1}{T^{1/2}}\sum_{t=1}^T \frac{t}{T}u_t\right) = \frac{T(T+1)(2T+1)}{6T^3}\sigma_u^2 \to \frac{\sigma_u^2}{3}.$$

It follows from Lemma 5.36 that

$$\frac{\sqrt{3}}{T^{1/2}\sigma_u}\sum_{t=1}^T \frac{t}{T}u_t \xrightarrow{D} \mathcal{N}(0,1).$$

These results show that $\{(t/T)u_t\}$ obeys a CLT, whereas $\{tu_t\}$ does not. □

Example 5.38 Suppose that y_t is generated as a random walk:

$$y_t = y_{t-1} + u_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, where u_t are i.i.d. random variables with mean zero and variance σ_u^2. From Example 5.31 we have seen that y_t have unbounded variances and strong correlations over time. Hence, they do not obey a CLT. Example 5.32 also suggests that $\{y_{t-1}u_t\}$ does not obey a CLT. □

In many applications, we usually encounter an array of functions of random variables and would like to know if it obeys a CLT. Let $\{z_{Tt}\}$ denote a triangular array of functions of random variables. Establishing a CLT may not be too difficult when $\{z_{Tt}\}$ is determined by sequences of independent random variables, but it is technically more involved when $\{z_{Tt}\}$ depends on sequences of correlated (or weakly dependent) random variables. In what follows, the array of square integrable random variables z_{Tt} is said to obey a CLT if

$$\frac{1}{\sigma_o\sqrt{T}}\sum_{t=1}^T [z_{Tt} - \mathbb{E}(z_{Tt})] = \frac{\sqrt{T}(\bar{z}_T - \bar{\mu}_T)}{\sigma_o} \xrightarrow{D} \mathcal{N}(0,1), \tag{5.3}$$

where $\bar{z}_T = T^{-1}\sum_{t=1}^T z_{Tt}$, $\bar{\mu}_T = \mathbb{E}(\bar{z}_T)$, and

$$\sigma_T^2 = \text{var}\left(T^{-1/2}\sum_{t=1}^T z_{Tt}\right) \to \sigma_o^2 > 0.$$

Note that this definition requires neither $\mathbb{E}(z_{Tt})$ nor $\text{var}(z_{Tt})$ to be a constant. If $\mathbb{E}(z_{Tt})$ is the constant μ_o, (5.3) would read:

$$\frac{\sqrt{T}(\bar{z}_T - \mu_o)}{\sigma_o} \xrightarrow{D} \mathcal{N}(0,1),$$

as we usually seen in other textbooks.

Consider an array of square integrable random vectors $\boldsymbol{z}_{Tt}$ in $\mathbb{R}^d$. Let $\bar{\boldsymbol{z}}_T$ denote the average of $\boldsymbol{z}_{Tt}$, $\bar{\boldsymbol{\mu}}_T = \mathbb{E}(\bar{\boldsymbol{z}}_T)$, and

$$\boldsymbol{\Sigma}_T = \text{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{z}_{Tt}\right) \to \boldsymbol{\Sigma}_o,$$

a positive definite matrix. Using the Cramér-Wold device (Lemma 5.18), $\{\boldsymbol{z}_{Tt}\}$ is said to obey a multivariate CLT, in the sense that

$$\boldsymbol{\Sigma}_o^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T[\boldsymbol{z}_{Tt} - \mathbb{E}(\boldsymbol{z}_{Tt})] = \boldsymbol{\Sigma}_o^{-1/2}\sqrt{T}(\bar{\boldsymbol{z}}_T - \bar{\boldsymbol{\mu}}_T) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d),$$

if $\{\boldsymbol{\alpha}'\boldsymbol{z}_{Tt}\}$ obeys a CLT, for any $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$.

5.8 Functional Central Limit Theorem

In this section, we consider a generalization of the concept of random variables and discuss the related limit theorem.

5.8.1 Stochastic Processes

Let $\mathcal{T}$ be a nonempty set of $\mathbb{R}$ and $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which the $\mathbb{R}^d$-valued random variables $\boldsymbol{z}_t$, $t \in \mathcal{T}$, are defined. Also let $(\mathbb{R}^d)^{\mathcal{T}}$ denote the collection of all $\mathbb{R}^d$-valued functions on $\mathcal{T}$, which is also a product space of copies of $\mathbb{R}^d$. For example, when $d = 1$ and $T = \{1, \ldots, k\}$, a real function on $\mathcal{T}$ is just a k-tuple $(z_1, \ldots, z_k)$, i.e., $(\mathbb{R})^{\{1,\ldots,k\}} = \mathbb{R}^k$; when $d = 1$ and T is an interval, $(\mathbb{R})^{\mathcal{T}}$ contains all real functions on that interval. A d-dimensional *stochastic process* with the *index set* $\mathcal{T}$ is a measurable mapping $\boldsymbol{z}\colon \Omega \mapsto (\mathbb{R}^d)^{\mathcal{T}}$ such that

$$\boldsymbol{z}(\omega) = \{\boldsymbol{z}_t(\omega),\, t \in \mathcal{T}\}.$$

For each $t \in \mathcal{T}$, $\boldsymbol{z}_t(\cdot)$ is a $\mathbb{R}^d$-valued random variable; for each ω, $\boldsymbol{z}(\omega)$ is a *sample path* (realization) of $\boldsymbol{z}$, which is a $\mathbb{R}^d$-valued function on $\mathcal{T}$. Therefore, a stochastic process is understood as a collection of random variables or a random function on the index set. The *random sequence* encountered in the preceding sections is just a stochastic process whose index set is the set of integers.

In what follows, for the stochastic process $\boldsymbol{z}$, we will write $\boldsymbol{z}(t,\cdot)$ or simply $\boldsymbol{z}(t)$ in place of $\boldsymbol{z}_t(\cdot)$. Thus, $\boldsymbol{z}$ with a subscript (say, $\boldsymbol{z}_n$) denotes a process in a sequence of stochastic processes. To signify the index set $\mathcal{T}$, we may also write $\boldsymbol{z}$ as $\{\boldsymbol{z}(t,\cdot),\, t \in \mathcal{T}\}$. The *finite-dimensional distributions* of $\{\boldsymbol{z}(t,\cdot),\, t \in \mathcal{T}\}$ is

$$\mathbb{P}(\boldsymbol{z}_{t_1} \le \boldsymbol{a}_1, \ldots, \boldsymbol{z}_{t_n} \le \boldsymbol{a}_n) = F_{t_1,\ldots,t_n}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n),$$

where $\{t_1, \ldots, t_n\}$ is any subset of $\mathcal{T}$ and $\boldsymbol{a}_i \in \mathbb{R}^d$. A stochastic process is said to be *stationary* if its finite-dimensional distributions are invariant under index displacement:

$$F_{t_1+s,\ldots,t_n+s}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) = F_{t_1,\ldots,t_n}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n).$$

A stochastic process is said to be *Gaussian* if its finite-dimensional distributions are all (multivariate) normal distributions.

The process $\{w(t),\, t \in [0,\infty)\}$ is the standard *Wiener process* (also known as the standard *Brownian motion*) if it has continuous sample paths almost surely and satisfies the following properties.

(i) $\mathbb{P}\big(w(0) = 0\big) = 1$.

(ii) For $0 \le t_0 \le t_1 \le \cdots \le t_k$,

$$\mathbb{P}\big(w(t_i) - w(t_{i-1}) \in B_i,\ i \le k\big) = \textstyle\prod_{i \le k} \mathbb{P}\big(w(t_i) - w(t_{i-1}) \in B_i\big),$$

where B_i are Borel sets.

(iii) For $0 \le s < t$, $w(t) - w(s)$ is normally distributed with mean zero and variance $t - s$.

By (i), this process must start from the origin with probability one. The second property requires non-overlapping increments of w being independent. By the property (iii), every increment of w is normally distributed with variance depending on the time difference; in particular, $w(t)$ is normally distributed with mean zero and variance t. This implies that for $r \le t$,

$$\operatorname{cov}\big(w(r),\, w(t)\big) = \mathbb{E}\big[w(r)\big(w(t) - w(r)\big)\big] + \mathbb{E}\big[w(r)^2\big] = r,$$

where $\mathbb{E}[w(r)(w(t) - w(r))] = 0$ because of independent increments.

The d-dimensional, standard Wiener process $\boldsymbol{w}$ is the process consisting of d mutually independent, standard Wiener processes. Thus, $\boldsymbol{w}$ still starts from the origin with probbability one, has independent increments, and

$$\boldsymbol{w}(t) - \boldsymbol{w}(s) \sim \mathcal{N}(\mathbf{0},\ (t-s)\,\boldsymbol{I}_d).$$

In view of the preceding paragraph, we have the following results for $\boldsymbol{w}$.

Lemma 5.39 *Let $\boldsymbol{w}$ be the d-dimensional, standard Wiener process.*

(i) $\boldsymbol{w}(t) \sim \mathcal{N}(\mathbf{0},\ t\,\boldsymbol{I}_d)$.

(ii) $\operatorname{cov}(\boldsymbol{w}(r),\ \boldsymbol{w}(t)) = \min(r,t)\,\boldsymbol{I}_d$.

We also note that, although the sample paths of the Wiener process are a.s. continuous, they are highly irregular. To see this, define $w_c(t) = w(c^2 t)/c$ for $c > 0$. It can be shown that w_c is also a standard Wiener process (Exercise 5.13). Note that $w_c(1/c) = w(c)/c$, where $w(c)/c$ is the slope of the chord between $w(c)$ and $w(0)$. If we choose a c large enough such that $w(c)/c > 1$, then the slope of the chord between $w_c(1/c)$ and $w_c(0)$ is

$$\frac{w_c(1/c)}{1/c} = \frac{w(c)/c}{1/c} = w(c) > c.$$

This shows that the sample path of w_c has a large slope c and hence must experience a large change on a very small interval $(0, 1/c)$. In fact, it can be shown that almost all the sample paths of w are *nowhere differentiable*; see e.g., Billingsley (1979, p. 450). Intuitively, the difference quotient $[w(t+h) - w(t)]/h$ is distributed as $\mathcal{N}(0,\ 1/|h|)$. As its variance diverges to infinity when h tends to zero, the difference quotient can not converge to a finite limit with a positive probability.

We may also construct different processes using the standard Wiener process. In particular, the process $\boldsymbol{w}^0$ on $[0,1]$ with $\boldsymbol{w}^0(t) = \boldsymbol{w}(t) - t\boldsymbol{w}(1)$ is known as the *Brownian bridge* or the *tied down Brownian motion*. It is easily seen that $\boldsymbol{w}^0(0) = \boldsymbol{w}^0(1) = 0$ with probability one so that the Brownian bridge starts from zero and must return to zero at $t = 1$. Moreover, $\mathbb{E}[\boldsymbol{w}^0(t)] = \mathbf{0}$, and for $r < t$,

$$\begin{aligned}\operatorname{cov}\big(\boldsymbol{w}^0(r),\ \boldsymbol{w}^0(t)\big) &= \operatorname{cov}\big(\boldsymbol{w}(r) - r\boldsymbol{w}(1),\ \boldsymbol{w}(t) - t\boldsymbol{w}(1)\big) \\ &= r(1-t)\,\boldsymbol{I}_d;\end{aligned}$$

in particular, $\operatorname{var}(\boldsymbol{w}^0(t)) = t(1-t)\,\boldsymbol{I}_d$ which reaches the maximum at $t = 1/2$.

5.8.2 Weak Convergence

Let S be a metric space and $\mathcal{S}$ be the Borel σ-algebra generated by the open sets in S. If for every bounded, continuous real function f on S we have

$$\int f(s)\,\mathrm{d}\mathbb{P}_n(s) \to \int f(s)\,\mathrm{d}\,\mathbb{P}(s),$$

where $\{\mathbb{P}_n\}$ and $\mathbb{P}$ are probability measures on $(S,\mathcal{S})$, we say that $\mathbb{P}_n$ *converges weakly* to $\mathbb{P}$ and write $\mathbb{P}_n \Rightarrow \mathbb{P}$. For the random elements $\boldsymbol{z}_n$ and $\boldsymbol{z}$ in S with the distributions induced by $\mathbb{P}_n$ and $\mathbb{P}$, respectively, we say that $\{\boldsymbol{z}_n\}$ *converges in distribution* to $\boldsymbol{z}$, also denoted as $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$, if $\mathbb{P}_n \Rightarrow \mathbb{P}$. Note that $\boldsymbol{z}_n$ and $\boldsymbol{z}$ here may be random functions. When $\boldsymbol{z}_n$ and $\boldsymbol{z}$ are all $\mathbb{R}^d$-valued random variables, $\mathbb{P}_n \Rightarrow \mathbb{P}$ reduces to the usual notion of convergence in distribution, as in Section 5.3.3. When $\boldsymbol{z}_n$ and $\boldsymbol{z}$ are d-dimensional stochastic processes, $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ implies that all the finite-dimensional distributions of $\boldsymbol{z}_n$ converge to the corresponding distributions of $\boldsymbol{z}$. To distinguish between the convergence in distribution of random variables and that of random functions, we shall, in what follows, denote the latter as $\boldsymbol{z}_n \Rightarrow \boldsymbol{z}$.

Let S and S' be two metric spaces with respective Borel σ-algebras $\mathcal{S}$ and $\mathcal{S}'$. Also let $g\colon S \mapsto S'$ be a measurable mapping. Then each probability measure $\mathbb{P}$ on $(S,\mathcal{S})$ induces a unique probability measure $\mathbb{P}^*$ on $(S',\mathcal{S}')$ via

$$\mathbb{P}^*(A') = \mathbb{P}(g^{-1}(A')), \qquad A' \in \mathcal{S}'.$$

If g is continuous almost everywhere on S, then for every bounded, continuous f on S', $f \circ g$ is also bounded and continuous on S. $\mathbb{P}_n \Rightarrow \mathbb{P}$ now implies that

$$\int f \circ g(s)\,\mathrm{d}\mathbb{P}_n(s) \to \int f \circ g(s)\,\mathrm{d}\,\mathbb{P}(s),$$

which is equivalent to

$$\int f(a)\,\mathrm{d}\mathbb{P}_n^*(a) \to \int f(a)\,\mathrm{d}\mathbb{P}^*(a).$$

This proves that $\mathbb{P}_n^* \Rightarrow \mathbb{P}^*$. This result is also known as the continuous mapping theorem; cf. Lemma 5.20.

Lemma 5.40 (Continuous Mapping Theorem) *Let $g\colon \mathbb{R}^d \mapsto \mathbb{R}$ be a function continuous almost everywhere on $\mathbb{R}^d$, except for at most countably many points. If $\boldsymbol{z}_n \Rightarrow \boldsymbol{z}$, then $g(\boldsymbol{z}_n) \Rightarrow g(\boldsymbol{z})$.*

For example, when $z_n \Rightarrow z$ and $h(z) = \sup_{0\le t\le 1} z(t)$,

$$\sup_{0\le t\le 1} z_n(t) \Rightarrow \sup_{0\le t\le 1} z(t),$$

and when $h(x) = \int_0^1 z(t)\,\mathrm{d}t$,

$$\int_0^1 z_n(t)\,\mathrm{d}t \Rightarrow \int_0^1 z(t)\,\mathrm{d}t.$$

5.8.3 Functional Central Limit Theorem

A sequence of random variables $\{\zeta_i\}$ is said to obey a *functional central limit theorem* (FCLT) if its normalized partial sums z_n converge in distribution to the standard Wiener process w, i.e., $z_n \Rightarrow w$. The FCLT, also known as the *invariance principle*, ensures that the limiting behavior of the normalized partial sums of ζ_i is governed by the standard Wiener process, regardless of the original distributions of ζ_i.

To see how the FCLT works, we consider the i.i.d. sequence $\{\zeta_i\}$ with mean zero and variance σ^2. The partial sum of ζ_i is $s_n = \zeta_1 + \cdots + \zeta_n$, and it can be normalized as $z_n(i/n) = (\sigma\sqrt{n})^{-1} s_i$. For $t \in [(i-1)/n, i/n)$, define the constant interpolations of $z_n(i/n)$ as

$$z_n(t) = z_n((i-1)/n) = \frac{1}{\sigma\sqrt{n}}\, s_{[nt]},$$

where $[nt]$ is the the largest integer less than or equal to nt, so that $[nt] = i - 1$. It can be seen that the sample paths of z_n are right continuous with left-hand limits, i.e., $z_n(t+) = z_n(t)$ and $z_n(t-) = \lim_{r \uparrow t} z_n(r)$. Such sample paths are also known as cadlag (an abreviation of the French "continue à droite, limite à gauche") functions. The interpolated process z_n is thus a random element of $D[0,1]$, the space of all cadlag functions. In view of the discussion of Section 5.8.2, we may study the weak convergence property of $\{z_n\}$.

We shall only discuss convergence of the finite-dimensional distributions of z_n. First note that, as n tends to infinity, we have from Lindeberg-Lévy's CLT that

$$\frac{1}{\sigma\sqrt{n}}\, s_{[nt]} = \left(\frac{[nt]}{n}\right)^{1/2} \frac{1}{\sigma\sqrt{[nt]}}\, s_{[nt]} \xrightarrow{D} \sqrt{t}\,\mathcal{N}(0,\,1),$$

which is just $\mathcal{N}(0,\,t)$, the distribution of $w(t)$. That is, $z_n(t) \xrightarrow{D} w(t)$. For $r < t$, we have

$$(z_n(r),\, z_n(t) - z_n(r)) \xrightarrow{D} \big(w(r),\, w(t) - w(r)\big),$$

from which we deduce that $(z_n(r),\, z_n(t)) \xrightarrow{D} (w(r),\, w(t))$. Proceeding along the same line we can show that all the finite-dimensional distributions of z_n converge to the corresponding distributions of the standard Wiener process. Although merely proving convergence of finite-dimensional distributions is not sufficient for $z_n \Rightarrow w$, it should help understanding the intuition of the FCLT. To arrive at $z_n \Rightarrow w$, it is also required the probability measures induced by z_n being "well behaved;" we omit the details.

In view of the discussion above, we are now ready to state an FCLT for i.i.d. random variables.

Lemma 5.41 (Donsker) *Let ζ_t be i.i.d. random variables with mean μ_o and variance $\sigma_o^2 > 0$ and z_T be the stochastic process with*

$$z_T(r) = \frac{1}{\sigma_o\sqrt{T}} \sum_{t=1}^{[Tr]} (\zeta_t - \mu_o), \quad r \in [0,1].$$

Then, $z_T \Rightarrow w$ as $T \to \infty$.

We observe from Lemma 5.41 that, when $r = 1$,

$$z_T(1) = \frac{\sqrt{T}(\bar{\zeta}_T - \mu_o)}{\sigma_o} \xrightarrow{D} \mathcal{N}(0,\, 1),$$

where $\bar{\zeta}_T = \sum_{t=1}^T \zeta_t/T$. This is precisely the conclusion of Lemma 5.35 and shows that Donsker's FCLT can be viewed as a generalization of Lindeberg-Lévy's CLT. The FCLT below applies to independent random variables and is a generalization of Liapunov's CLT (Lemma 5.36); see White (2001).

Lemma 5.42 *Let ζ_t be independent random variables with mean μ_t and variance $\sigma_t^2 > 0$ such that*

$$\bar{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^T \sigma_t^2 \to \sigma_o^2 > 0.$$

Also let z_T be the stochastic process with

$$z_T(r) = \frac{1}{\sigma_o\sqrt{T}} \sum_{t=1}^{[Tr]} (\zeta_t - \mu_t), \quad r \in [0,1].$$

If for some $\delta > 0$, $\mathbb{E}\,|\zeta_t|^{2+\delta}$ are bounded for all t, then $z_T \Rightarrow w$ as $T \to \infty$.

More generally, let ζ_t be (possibly dependent and heterogeneously distributed) random variables with mean μ_t and variance $\sigma_t^2 > 0$. Define the *long-run variance* of ζ_t as

$$\sigma_*^2 = \lim_{T\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \zeta_t\right),$$

and assume σ_*^2 exists and is positive. We say that $\{\zeta_t\}$ obeys an FCLT if $z_T \Rightarrow w$ as $T \to \infty$, where z_T is the stochastic process with

$$z_T(r) = \frac{1}{\sigma_*\sqrt{T}} \sum_{t=1}^{[Tr]} (\zeta_t - \mu_t), \quad r \in [0,1].$$

When ζ_t are independent random variables, $\operatorname{cov}(\zeta_t, \zeta_s) = 0$ for all $t \neq s$, so that $\sigma_*^2 = \sigma_o^2$. Then the generic FCLT above leads to the conclusion of Lemma 5.42.

Let $\boldsymbol{\zeta}_t$ are d-dimensional random variables with mean $\boldsymbol{\mu}_t$ and variance-covariance matrices $\boldsymbol{\Sigma}_t^2$. Define the *long-run variance-covariance matrix* of $\boldsymbol{\zeta}_t$ as

$$\boldsymbol{\Sigma}_* = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\left(\sum_{t=1}^{T}(\boldsymbol{\zeta}_t - \boldsymbol{\mu}_t)\right)\left(\sum_{t=1}^{T}(\boldsymbol{\zeta}_t - \boldsymbol{\mu}_t)\right)'\right],$$

and assume that $\boldsymbol{\Sigma}_*$ exists and is positive definite. We say that $\{\boldsymbol{\zeta}_t\}$ obeys a (multivariate) FCLT if $\boldsymbol{z}_T \Rightarrow \boldsymbol{w}$ as $T \to \infty$, where $\boldsymbol{z}_T$ is the d-dimensional stochastic process with

$$\boldsymbol{z}_T(r) = \frac{1}{\sqrt{T}} \boldsymbol{\Sigma}_*^{-1/2} \sum_{t=1}^{[Tr]} (\boldsymbol{\zeta}_t - \boldsymbol{\mu}_t), \quad r \in [0,1],$$

and $\boldsymbol{w}$ is the d-dimensional, standard Wiener process. Although no sufficient conditions will be provided, we note that a FCLT may hold for weakly dependent and heterogeneously distributed data, provided that they satisfy some regularity conditions; see Davidson (1994) and White (2001) for details.

Example 5.43 Suppose that y_t is generated as a random walk:

$$y_t = y_{t-1} + u_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, where u_t are i.i.d. random variables with mean zero and variance σ_u^2. As $\{u_t\}$ obeys Donsker's FCLT and $y_{[Tr]} = \sum_{t=1}^{[Tr]} u_t$ is a partial sum of u_t, we have

$$\begin{aligned}\frac{1}{T^{3/2}} \sum_{t=1}^{T} y_t &= \sigma_u \sum_{t=1}^{T} \int_{(t-1)/T}^{t/T} \frac{1}{\sqrt{T}\sigma_u} y_{[Tr]} \, \mathrm{d}r \\ &\Rightarrow \sigma_u \int_0^1 w(r) \, \mathrm{d}r,\end{aligned}$$

where the right-hand side is a random variable. This result also verifies that $\sum_{t=1}^{T} y_t$ is $O_{\mathbb{P}}(T^{3/2})$, as stated in Example 5.31. Similarly,

$$\frac{1}{T^2} \sum_{t=1}^{T} y_t^2 \Rightarrow \sigma_u^2 \int_0^1 w(r)^2 \, \mathrm{d}r,$$

so that $\sum_{t=1}^{T} y_t^2$ is $O_{\mathbb{P}}(T^2)$. It is clear that these results remain valid, as long as u_t obey a FCLT (but need not be i.i.d. or independent). □

Exercises

5.1 Let $\mathcal{C}$ be a collection of subsets of Ω. Show that the intersection of all the σ-algebras on Ω that contain $\mathcal{C}$ is the smallest σ-algebra containing $\mathcal{C}$.

5.2 Let y and z be two independent, integrable random variables. Show that $\mathbb{E}(yz) = \mathbb{E}(y)\,\mathbb{E}(z)$.

5.3 Let x and y be two random variables with finite p^{th} moment ($p > 1$). Prove the following triangle inequality:

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Hint: Write $\mathbb{E}\,|x + y|^p = \mathbb{E}(|x + y||x + y|^{p-1})$ and apply Hölder's inequality.

5.4 In the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ suppose that we know the event B in $\mathcal{F}$ has occurred. Show that the conditional probability $\mathbb{P}(\cdot|B)$ satisfies the axioms for probability measures.

5.5 Prove the first assertion of Lemma 5.9.

5.6 Prove that for the square integrable random vectors $\boldsymbol{z}$ and $\boldsymbol{y}$,

$$\text{var}(\boldsymbol{z}) = \mathbb{E}[\text{var}(\boldsymbol{z} \mid \boldsymbol{y})] + \text{var}(\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})).$$

5.7 A sequence of square integrable random variables $\{z_n\}$ is said to converge to a random variable z in L_2 (in quadratic mean) if

$$\mathbb{E}(z_n - z)^2 \to 0.$$

Prove that L_2 convergence implies convergence in probability.
Hint: Apply Chebychev's inequality.

5.8 Show that a sequence of square integrable random variables $\{z_n\}$ converges to a constant c in L_2 if and only if $\mathbb{E}(z_n) \to c$ and $\text{var}(z_n) \to 0$.

5.9 Prove that $\boldsymbol{z}_T \overset{A}{\sim} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I})$ if, and only if, $\boldsymbol{\lambda}'\boldsymbol{z}_T \overset{A}{\sim} \mathcal{N}(0, 1)$ for all $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$.

5.10 Prove Lemma 5.23.

5.11 Suppose that $\mathbb{E}(z_n^2) = O(c_n)$, where $\{c_n\}$ is a sequence of positive real numbers. Show that $z_n = O_{\mathbb{P}}(c_n^{1/2})$.

5.12 Suppose that y_t is generated as a Gaussian random walk:

$$y_t = y_{t-1} + u_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, where u_t are i.i.d. normal random variables with mean zero and variance σ_u^2. Show that $\sum_{t=1}^T y_t^2$ is $O_{\mathbb{P}}(T^2)$.

5.13 Let w be a standard Wierner process and define w_c as $w_c(t) = w(c^2 t)/c$, where $c > 0$. Show that w_c is also a standard Wierner process.

5.14 Let w be a standard Wiener process and w^0 a Brownian bridge. Suppose that $x(t) = w(t + r) - w(r)$ for a given $r > 0$ and $y(t) = (1 + t)\, w^0(t/(1 + t))$, $t \in [0, \infty)$. Show that both x and y are standard Wiener processes.

References

Ash, Robert B. (1972). *Real Analysis and Probability*, New York, NY: Academic Press.

Bierens, Herman J. (1994). *Topics in Advanced Econometrics*, New York, NY: Cambridge University Press.

Billingsley, Patrick (1979). *Probability and Measure*, New York, NY: John Wiley and Sons.

Davidson, James (1994). *Stochastic Limit Theory*, New York, NY: Oxford University Press.

Gallant, A. Ronald (1997). *An Introduction to Econometric Theory*, Princeton, NJ: Princeton University Press.

Gallant, A. Ronald and Halbert White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford, UK: Basil Blackwell.

White, Halbert (2001). *Asymptotic Theory for Econometricians*, revised edition, Orlando, FL: Academic Press.