2

Modes of Convergence

2.1 Introduction

In this chapter we discuss different types of convergence in probability and statistics. Types of convergence have already been introduced. They are convergence in probability and convergence in distribution. In addition, we introduce other types of convergence, such as almost sure convergence and $L^p$ convergence. We discuss properties of different types of convergence, the connections between them, and how to establish these properties. The discussion will mainly focus on the case of univariate random variables. However, most of the results presented here can be easily extended to the multivariate situation.

The concept of different types of convergence is critically important in mathematical statistics. In fact, misusage of such concepts often leads to confusions, even errors. The following is a simple example.

Example 2.1 (Asymptotic variance). The well-known result of CLT states that, under regularity conditions, we have $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$, where $\sigma^2$ is called the asymptotic variance. The definition seems to be clear enough: $\sigma^2$ is the variance of the limiting normal distribution. Even so, some confusion still arises, and the following are some of them.

(a) $\sigma^2$ is the asymptotic variance of $\bar{X}$.
(b) $\lim_{n \to \infty} n \text{var}(\bar{X}) \to \sigma^2$ as $n \to \infty$.
(c) $n(\bar{X} - \mu)^2 \to \sigma^2$ as $n \to \infty$.

Statement (a) is clearly incorrect. It would be more appropriate to say that $\sigma^2$ is the asymptotic variance of $\sqrt{n}\bar{X}$; however, this does not mean that $\lim_{n \to \infty} \text{var}(\sqrt{n}\bar{X}) \to \sigma^2$, as $n \to \infty$, or Statement (b). In fact, convergence in distribution and convergence of the variance (which is essentially the convergence of moments) are two different concepts, and they do not imply each other. In some cases, even if the variance does not exist, the CLT still holds (e.g., Ibragimov and Linnik 1971, pp. 85, Theorem 2.6.3). As for Statement (c), it is not clear in what sense the convergence is. Even if the latter is made
clear, say, in probability, it is still a wrong statement because, according to the CLT, \( n(\bar{X} - \mu)^2 = \left( \sqrt{n}(X - \mu) \right)^2 \) converges in distribution to \( \sigma^2 \chi^2_1 \), where \( \chi^2_1 \) is the \( \chi^2 \)-distribution with one degree of freedom. Since the latter is a random variable, not a constant, Statement (c) is incorrect even in the sense of convergence in probability.

In a way, the problem associated with Statement (c) is very similar to the second example in the Preface regarding approximation to a mean.

2.2 Convergence in probability

For the sake of completeness, here is the definition once again. A sequence of random variables \( \xi_1, \xi_2, \ldots \) converges in probability to a random variable \( \xi \), denoted by \( \xi_n \xrightarrow{P} \xi \), if for any \( \epsilon > 0 \), we have \( P(\left| \xi_n - \xi \right| > \epsilon) \to 0 \) as \( n \to \infty \).

It should be pointed out that, more precisely, convergence in probability is a property about the distributions of the random variables \( \xi_1, \xi_2, \ldots, \xi \) rather than the random variables themselves. In particular, convergence in probability does not imply that the sequence \( \xi_1, \xi_2, \ldots \) converges pointwisely at all. For example, consider the following.

**Example 2.2.** Define the sequence of random variables \( \xi_n = \xi_n(x) \), \( x \in [0, 1] \), which is the probability space with the probability being the Lebesgue measure (see Appendix A.2), as follows.

\[
\begin{align*}
\xi_1(x) &= \begin{cases} 
1, & x \in [0, 1/2) \\
0, & x \in [1/2, 1] 
\end{cases} \\
\xi_2(x) &= \begin{cases} 
0, & x \in [0, 1/2) \\
1, & x \in [1/2, 1] 
\end{cases} \\
\xi_3(x) &= \begin{cases} 
1, & x \in [0, 1/4) \\
0, & x \in [0, 1) \setminus [0, 1/4) 
\end{cases} \\
\xi_4(x) &= \begin{cases} 
1, & x \in [1/4, 1/2) \\
0, & x \in [0, 1) \setminus [1/4, 1/2) 
\end{cases} \\
\xi_5(x) &= \begin{cases} 
1, & x \in [1/2, 3/4) \\
0, & x \in [0, 1) \setminus [1/2, 3/4) 
\end{cases} \\
\xi_6(x) &= \begin{cases} 
1, & x \in [3/4, 1) \\
0, & x \in [0, 1) \setminus [3/4, 1] 
\end{cases}
\end{align*}
\]

and so forth (see Figure 2.1). It can be shown that \( \xi_n \xrightarrow{P} 0 \) as \( n \to \infty \); however, \( \xi_n(x) \) does not converge pointwisely at any \( x \in [0, 1] \) (Exercise 2.1).

So, what does convergence in probability really mean after all? It means that the overall probability that \( \xi_n \) is not close to \( \xi \) goes to zero as \( n \) increases, and nothing more than that. We consider another example.
Example 2.3. Suppose that $\xi_n$ is uniformly distributed over the intervals

$$\left[ i - \frac{1}{2n^2}, i + \frac{1}{2n^2} \right], \quad i = 1, \ldots, n.$$ 

Then the sequence $\xi_n$, $n \geq 1$, converges in probability to zero. To see this, note that the pdf of $\xi_n$ is given by

$$f_n(x) = \begin{cases} n, & x \in [i - 1/2n^2, i + 1/2n^2], \quad 1 \leq i \leq n \\ 0, & \text{elsewhere}. \end{cases}$$

It follows that for any $\epsilon > 0$, $P(|\xi_n| > \epsilon) = 1/n \to 0$, as $n \to \infty$; hence, $\xi_n \xrightarrow{P} 0$. The striking thing about this example is that, as $n \to \infty$, the height of the density function actually approaches infinity. Meanwhile, the total area in which the density is nonzero approaches zero as $n \to \infty$, which is what counts in the convergence in probability of the sequence (see Figure 2.2).
The follow theorems provide useful sufficient conditions for convergence in probability.

**Theorem 2.1.** Suppose that $E(|ξ_n - ξ|^p) → 0$ as $n → ∞$ for some $p > 0$. Then $ξ_n →^P ξ$ as $n → ∞$.

The proof follows from the Chebyshev’s inequality (Exercise 2.2).

**Theorem 2.2.** Suppose that $ξ_n = a_n η_n + b_n$, where $a_n$ and $b_n$ are non-random sequences such that $a_n → a$, $b_n → b$ as $n → ∞$, and $η_n$ is a sequence of random variables such that $η_n →^P η$ as $n → ∞$. Then $ξ_n →^P ξ = aη + b$ as $n → ∞$. 

*Fig. 2.2. A plot of the pdfs of the random variables in Example 2.3*
Theorem 2.3. Suppose that $\xi_n \xrightarrow{P} \xi$ and $\eta_n \xrightarrow{P} \eta$ as $n \to \infty$. Then $\xi_n + \eta_n \xrightarrow{P} \xi + \eta$ as $n \to \infty$.

Theorem 2.4. Suppose that $\xi_n \xrightarrow{P} \xi$ and $\xi$ is positive with probability 1. Then $\xi_n^{-1} \xrightarrow{P} \xi^{-1}$ as $n \to \infty$.

The proofs of Theorems 2.2–2.4 are left to the readers as exercises (Exercises 2.3–2.5).

An important property of convergence in probability is the following. The sequence $\xi_n, n = 1, 2, \ldots$, is called bounded in probability if for any $\epsilon > 0$, there is $M > 0$ such that $P(|\xi_n| \leq M) > 1 - \epsilon$ for any $n \geq 1$.

Theorem 2.5. If $\xi_n, n = 1, 2, \ldots$, converges in probability, then the sequence is bounded in probability.

Proof. Suppose that $\xi_n \xrightarrow{P} \xi$ for some random variable $\xi$. Then for any $\epsilon > 0$, there is $B > 0$ such that $P(|\xi| \leq B) > 1 - \epsilon$ (see Example A.5). On the other hand, by convergence in probability, there is $N \geq 1$ such that when $n \geq N$, we have $P(|\xi_n - \xi| \leq 1) > 1 - \epsilon$. It follows that

$$P(|\xi_n| \leq B + 1) \geq P(|\xi_n - \xi| \leq 1, |\xi| \leq B) > 1 - 2\epsilon, \quad n \geq N.$$ 

Now, let $\eta$ be the random variable $\max_{1 \leq n \leq N-1} |\xi_n|$. According to Example A.5, there is a constant $A > 0$ such that $P(\eta \leq A) > 1 - 2\epsilon$. Let $M = A \lor (B + 1)$. Then we have $P(|\xi_n| \leq M) > 1 - 2\epsilon, n \geq 1$. Since $\epsilon$ is arbitrary, this completes the proof. Q.E.D.

With the help of Theorem 2.5 it is easy to establish the following result (Exercise 2.6).

Theorem 2.6. Suppose that $\xi_n \xrightarrow{P} \xi$ and $\eta_n \xrightarrow{P} \eta$ as $n \to \infty$. Then $\xi_n \eta_n \xrightarrow{P} \xi \eta$ as $n \to \infty$.

2.3 Almost sure convergence

A sequence of random variables $\xi_n, n = 1, 2, \ldots$, converges almost surely to a random variable $\xi$, denoted by $\xi_n \xrightarrow{a.s.} \xi$ if $P(\lim_{n \to \infty} \xi_n = \xi) = 1$.

Almost sure convergence is a stronger property than convergence in probability, as the following theorem shows.

Theorem 2.7. $\xi_n \xrightarrow{a.s.} \xi$ implies $\xi_n \xrightarrow{P} \xi$. 

The proof follows from the following lemma whose proof is a good exercise of the $\epsilon$-$\delta$ argument discussed in Chapter 1 (Exercise 2.11).

**Lemma 2.1.** $\xi_n \overset{\text{a.s.}}{\rightarrow} \xi$ if and only if for every $\epsilon > 0$,
\[
\lim_{N \rightarrow \infty} P(\cup_{n=N}^{\infty} \{||\xi_n - \xi|| \geq \epsilon\}) = 0. \tag{2.1}
\]

On the other hand, Example 2.2 shows that there are sequences of random variables that converge in probability but not almost surely. We consider some more examples.

**Example 2.4.** Consider the same probability space $[0, 1]$ as in Example 2.2 but a different sequence of random variables $\xi_n, n = 1, 2, \ldots$, defined as follows: $\xi_n(i/n) = i, 1 \leq i \leq n$, and $\xi_n(x) = 0$, if $x \in [0, 1] \setminus \{i/n, 1 \leq i \leq n\}$. Then $\xi_n \overset{\text{a.s.}}{\rightarrow} 0$ as $n \rightarrow \infty$. To see this, let $A = \{i/n, i = 1, \ldots, n, n = 1, 2, \ldots\}$. Then $P(A) = 0$ (note that $P$ is the Lebesgue measure on $[0, 1]$). Furthermore, for any $x \in [0, 1] \setminus A$, we have $\xi_n(x) = 0$ for any $n$; hence, $\xi_n(x) \rightarrow 0$ as $n \rightarrow \infty$. Therefore, $P(\lim_{n \rightarrow \infty} \xi_n = 0) \geq P([0, 1] \setminus A) = 1$.

**Example 2.5.** Suppose that $X_i$ is a random variable with a Binomial$(i, p)$ distribution, $i = 1, 2, \ldots$, where $p \in [0, 1]$. Define
\[
\xi_n = \sum_{i=1}^{n} \frac{X_i}{i^{2+\delta}}, \quad n = 1, 2, \ldots,
\]
where $\delta > 0$. Then
\[
\xi_n \overset{\text{a.s.}}{\rightarrow} \xi = \sum_{i=1}^{\infty} \frac{X_i}{i^{2+\delta}} \text{ as } n \rightarrow \infty. \tag{2.2}
\]
To see this, note that $0 \leq X_i/i^{2+\delta} \leq i/i^{2+\delta} = 1/i^{1+\delta}$, and the infinite series $\sum_{i=1}^{\infty} 1/i^{1+\delta}$ converges. Therefore, by the result of §1.5.2.9 (i), the infinite series $\sum_{i=1}^{\infty} X_i/i^{2+\delta}$ always converges, which implies (2.2).

The following result is often useful in proving almost sure convergence.

**Theorem 2.8.** If, for every $\epsilon > 0$, we have $\sum_{n=1}^{\infty} P(||\xi_n - \xi|| \geq \epsilon) < \infty$, then $\xi_n \overset{\text{a.s.}}{\rightarrow} \xi$ as $n \rightarrow \infty$.

**Proof.** By Lemma 2.1 we need to show (2.1). Since
\[
P(\cup_{n=N}^{\infty} \{||\xi_n - \xi|| \geq \epsilon\}) \leq \sum_{n=N}^{\infty} P(||\xi_n - \xi|| \geq \epsilon),
\]
and the latter converges to zero as $N \rightarrow \infty$, because the sequence $\sum_{n=1}^{\infty} P(||\xi_n - \xi|| \geq \epsilon)$ is convergent, the result follows. Q.E.D.
Example 2.6. In Example 1.4 we showed consistency of the MLE in the Uniform distribution, that is, $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$, where $\hat{\theta} = X_{(n)}$ and $X_1, \ldots, X_n$ are i.i.d. observations from the Uniform[0, $\theta$] distribution. We now show that, in fact, $\hat{\theta} \xrightarrow{a.s.} \theta$ as $n \to \infty$. For any $\epsilon > 0$, we have

$$P\{|X_{(n)} - \theta| \geq \epsilon\} = P\{X_{(n)} \leq \theta - \epsilon\}$$

$$= P(\{X_1 \leq \theta - \epsilon, \ldots, X_n \leq \theta - \epsilon\}$$

$$= \{P(X_1 \leq \theta - \epsilon\})^n$$

$$= \left(1 - \frac{\epsilon}{\theta}\right)^n.$$

Thus, we have

$$\sum_{n=1}^{\infty} P\{|X_{(n)} - \theta| \geq \epsilon\} = \sum_{n=1}^{\infty} \left(1 - \frac{\epsilon}{\theta}\right)^n$$

$$= \frac{\theta - \epsilon}{\epsilon} < \infty.$$

Here, we assume, without loss of generality, that $\epsilon < \theta$. It follows by Theorem 2.8 that $X_{(n)} \xrightarrow{a.s.} \theta$ as $n \to \infty$.

The following example is known as the bounded strong law of large numbers, which is a special case of the strong law of large numbers (SLLN; see Chapter 6).

Example 2.7. Suppose that $X_1, \ldots, X_n$ are i.i.d. and $|X_i| \leq b$ for some constant $b$. Then

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} E(X_1)$$

as $n \to \infty$. To show (2.3), note that, for any $\epsilon > 0$,

$$P\{|\xi_n - E(X_1)| \geq \epsilon\} = P\left\{\frac{1}{n} \sum_{i=1}^{n} X_i - E(X_1) \geq \epsilon\right\}$$

$$+ P\left\{\frac{1}{n} \sum_{i=1}^{n} X_i - E(X_1) \leq -\epsilon\right\}$$

$$= I_1 + I_2. \quad (2.4)$$

Furthermore, we have, by Chebyshev’s inequality (see Section 5.2),

$$I_1 = P\left[\sum_{i=1}^{n} \left\{\frac{X_i - E(X_1)}{\sqrt{n}}\right\} \geq \epsilon\sqrt{n}\right]$$

$$= P\left(\exp\left[\sum_{i=1}^{n} \left\{\frac{X_i - E(X_1)}{\sqrt{n}}\right\}\right] \geq e^{\epsilon\sqrt{n}}\right)$$
2 Modes of Convergence

\[
\leq e^{-\epsilon\sqrt{n}} E \left( \exp \left[ \sum_{i=1}^{n} \left\{ \frac{X_i - E(X_1)}{\sqrt{n}} \right\} \right] \right) \\
= e^{-\epsilon\sqrt{n}} E \left[ \prod_{i=1}^{n} \exp \left\{ \frac{X_i - E(X_1)}{\sqrt{n}} \right\} \right] \\
= e^{-\epsilon\sqrt{n}} \left( E \left[ \exp \left\{ \frac{X_1 - E(X_1)}{\sqrt{n}} \right\} \right] \right)^n. \tag{2.5}
\]

By Taylor’s expansion (see Section 4.1), we have, for any \( x \in R \),

\[
e^x = 1 + x + \frac{e^{\lambda x}}{2} x^2
\]

for some \( 0 \leq \lambda \leq 1 \). It follows that \( e^x \leq 1 + x + (e^c/2)x^2 \) if \( |x| \leq c \). Since \( |X_1 - E(X_1)|/\sqrt{n} \leq 2b/\sqrt{n} \leq 2b \), by letting \( c = 2b \) we have

\[
\exp \left\{ \frac{X_1 - E(X_1)}{\sqrt{n}} \right\} \leq 1 + \frac{X_1 - E(X_1)}{\sqrt{n}} + e^{2b} \left\{ \frac{X_1 - E(X_1)}{\sqrt{n}} \right\}^2 \\
\leq 1 + \frac{X_1 - E(X_1)}{\sqrt{n}} + \frac{2b^2 e^{2b}}{n}
\]

(because \( |X_1 - E(X_1)| \leq 2b \)); hence,

\[
E \left[ \exp \left\{ \frac{X_1 - E(X_1)}{\sqrt{n}} \right\} \right] \leq 1 + \frac{2b^2 e^{2b}}{n} \\
\leq \exp \left( \frac{2b^2 e^{2b}}{n} \right) \tag{2.6}
\]

using the inequality \( e^x \geq 1 + x \) for all \( x \geq 0 \). By (2.5) and (2.6), we have \( I_1 \leq ce^{-\epsilon\sqrt{n}} \), where \( c = \exp(2b^2 e^{2b}) \). By similar arguments, it can be shown that \( I_2 \leq ce^{-\epsilon\sqrt{n}} \) (Exercise 2.12). Therefore, by (2.4), we have \( P(|\xi_n - E(X_1)| \geq \epsilon) \leq 2ce^{-\epsilon\sqrt{n}} \). The almost sure convergence of \( \xi_n \) to \( E(X_1) \) then follows from Theorem 1.8, because \( \sum_{i=1}^{\infty} e^{-\epsilon\sqrt{n}} < \infty \) (Exercise 2.13).

2.4 Convergence in distribution

Convergence in distribution is another concept that was introduced earlier. Again, for the sake of completeness we repeat the definition here. A sequence of random variables \( \xi_1, \xi_2, \ldots \) converges in distribution to a random variable \( \xi \), denoted by \( \xi_n \xrightarrow{d} \xi \), if \( F_n \xrightarrow{w} F \), where \( F_n \) is the cdf of \( \xi_n \) and \( F \) is the cdf of \( \xi \). The latter means that \( F_n(x) \to F(x) \) as \( n \to \infty \) for every \( x \) at which \( F(x) \) is continuous.

Note that convergence in distribution is a property of the distribution of \( \xi_n \) rather than \( \xi_n \) itself. In particular, convergence in distribution does not imply almost sure convergence or even convergence in probability.
Example 2.8. Let \( \xi \) be a random variable that has the standard normal distribution, \( N(0,1) \). Let \( \xi_1 = \xi, \xi_2 = -\xi, \xi_3 = \xi, \xi_4 = -\xi \), and so forth. Then, clearly, \( \xi_n \xrightarrow{d} \xi \) (because \( \xi \) and \( -\xi \) have the same distribution). On the other hand, \( \xi_n \) does not converge in probability to \( \xi \) or any other random variable \( \eta \). To see this, suppose that \( \xi_n \xrightarrow{P} \eta \) for some random variable \( \eta \). Then we must have \( P(|\xi_n - \eta| > 1) \rightarrow 0 \) as \( n \rightarrow \infty \). Therefore, we have

\[
\begin{align*}
P(|\xi - \eta| > 1) &= P(|\xi_{2k-1} - \eta| > 1) \rightarrow 0, \quad (2.7) \\
P(|\xi + \eta| > 1) &= P(|\xi_{2k} - \eta| > 1) \rightarrow 0 \quad (2.8)
\end{align*}
\]

as \( k \rightarrow \infty \). Because the left sides of (2.7) and (2.8) do not depend on \( k \), we must have \( P(|\xi - \eta| > 1) = 0 \) and \( P(|\xi + \eta| > 1) = 0 \). Then because \(|2\xi| \leq |\xi - \eta| + |\xi + \eta|, |\xi| > 1 \) implies \(|2\xi| > 2 \), which, in turn, implies that either \(|\xi - \eta| > 1 \) or \(|\xi + \eta| > 1 \). It follows that \( P(|\xi| > 1) \leq P(|\xi - \eta| > 1) + P(|\xi + \eta| > 1) = 0 \), which is, of course, not true.

Since almost sure convergence implies convergence in probability (Theorem 2.7), the sequence \( \xi_n \) in Example 2.8 also does not converge almost surely. Nevertheless, the fact that the distribution of \( \xi_n \) is the same for any \( n \) is enough to imply convergence in distribution.

On the other hand, the following theorem shows that convergence in probability indeed implies convergence in distribution, so the former is a stronger convergent property than the latter.

**Theorem 2.9.** \( \xi_n \xrightarrow{P} \xi \) implies \( \xi_n \xrightarrow{d} \xi \).

**Proof.** Let \( x \) be a continuity point of \( F(x) \). We need to show that \( P(\xi_n \leq x) = F_n(x) \rightarrow F(x) = P(\xi \leq x) \). For any \( \epsilon > 0 \), we have

\[
F(x - \epsilon) = P(\xi \leq x - \epsilon)
\]

\[
= P(\xi \leq x - \epsilon, \xi_n \leq x) + P(\xi \leq x - \epsilon, \xi_n > x)
\]

\[
\leq P(\xi_n \leq x) + P(|\xi_n - \xi| > \epsilon)
\]

\[
= F_n(x) + P(|\xi_n - \xi| > \epsilon).
\]

It follows by the results of §1.5.1.5 that

\[
F(x - \epsilon) \leq \liminf F_n(x) + \limsup P(|\xi_n - \xi| > \epsilon)
\]

\[
= \liminf F_n(x).
\]

By a similar argument, it can be shown that (Exercise 2.18)

\[
F(x + \epsilon) \geq \limsup F_n(x).
\]

Since \( \epsilon \) is arbitrary and \( F(x) \) is continuous at \( x \), we have

\[
\limsup F_n(x) \leq F(x) \leq \liminf F_n(x).
\]
On the other hand, we always have \( \lim \inf F_n(x) \leq \lim \sup F_n(x) \). Therefore, we have \( \lim \inf F_n(x) = \lim \sup F_n(x) = F(x) \); hence, \( \lim_{n \to \infty} F_n(x) = F(x) \) by the results of §1.5.1.2. This completes the proof. Q.E.D.

Although convergence in distribution can often be verified by the definition, the following theorems sometimes offer more convenient tools for establishing convergence in distribution.

Let \( \xi \) be a random variable. The moment generating function (mgf) of \( \xi \) is defined as
\[
m_\xi(t) = E(e^{t \xi}),
\] (2.9)
provided that the expectation exists; the characteristic function (cf) of \( \xi \) is defined as
\[
c_\xi(t) = E(e^{it \xi}),
\] (2.10)
where \( i = \sqrt{-1} \). Note that the mgf is defined at \( t \in \mathbb{R} \) for which the expectation (2.9) exists (i.e., finite). It is possible, however, that the expectation does not exist for any \( t \) except \( t = 0 \) (Exercise 2.19). The latter is the one particular value of \( t \) at which the mgf is always well defined. On the other hand, the cf is well defined for any \( t \in \mathbb{R} \). This is because \( |e^{it \xi}| \leq 1 \) by the properties of complex numbers (Exercise 2.20).

**Theorem 2.10.** Let \( m_n(t) \) be the mgf of \( \xi_n, n = 1, 2, \ldots \). Suppose that there is \( \delta > 0 \) such that \( m_n(t) \to m(t) \) as \( n \to \infty \) for all \( t \) such that \( |t| < \delta \), where \( m(t) \) is the mgf of a random variable \( \xi \); then \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \).

In other words, convergence of the mgf in a neighborhood of zero implies convergence in distribution. The following example shows that the converse of Theorem 2.10 is not true; that is, convergence in distribution does not necessarily imply convergence of the mgf in a neighborhood of zero.

**Example 2.9.** Suppose that \( \xi_n \) has a \( t \)-distribution with \( n \) degrees of freedom (i.e., \( \xi_n \sim t_n \)). Then it can be shown that \( \xi_n \xrightarrow{d} \xi \sim N(0, 1) \) as \( n \to \infty \). However, \( m_n(t) = E(e^{t \xi_n}) = \infty \) for any \( t \neq 0 \), whereas the mgf of \( \xi \) is given by \( m(t) = e^{t^2/2}, t \in \mathbb{R} \) (Exercise 2.21). Therefore, \( m_n(t) \) does not converge to \( m(t) \) for any \( t \neq 0 \).

On the other hand, convergence of the cf is indeed equivalent to convergence in distribution, as the following theorem shows.

**Theorem 2.11** (Lévy-Cramér continuity theorem). Let \( c_n(t) \) be the cf of \( \xi_n, n = 1, 2, \ldots \), and \( c(t) \) be the cf of \( \xi \). Then \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \) if and only if \( c_n(t) \to c(t) \) as \( n \to \infty \) for every \( t \in \mathbb{R} \).

The proof of Theorem 2.10 is based on the theory of Laplace transformation. Consider, for example, the case that \( \xi \) is a continuous random variable
that has the pdf $f_{\xi}(x)$ with respect to the Lebesgue measure (see Appendix A.2). Then

$$m_{\xi}(t) = \int_{-\infty}^{\infty} e^{tx} f_{\xi}(x) \, dx,$$

(2.11)

which is the Laplace transformation of $f(x)$. A nice property of the Laplace transformation is its uniqueness. This means that if (2.11) holds for all $t$ such that $|t| < \delta$, where $\delta > 0$, then there is one and only one $f_{\xi}(x)$ that satisfies (2.11). Given this property, it is not surprising that Theorem 2.10 holds, and this actually outlines the main idea of the proof. The idea behind the proof of Theorem 2.11 is similar. We omit the details of both proofs, which are technical in nature (e.g., Feller 1971).

The following properties of the mgf and cf are often useful. The proofs are left as exercises (Exercises 2.22, 2.23).

**Lemma 2.2.** (i) Let $\xi$ be a random variable. Then, for any constants $a$ and $b$, we have

$$m_{a\xi+b}(t) = e^{bt} m_{\xi}(at),$$

provided that the $m_{\xi}(at)$ exists. (ii) Let $\xi, \eta$ be independent random variables. Then we have

$$m_{\xi+\eta}(t) = m_{\xi}(t)m_{\eta}(t), \quad |t| \leq \delta,$$

provided that both $m_{\xi}(t)$ and $m_{\eta}(t)$ exist.

**Lemma 2.3.** (i) Let $\xi$ be a random variable. Then, for any constants $a$ and $b$, we have

$$c_{a\xi+b}(t) = e^{ibt} c_{\xi}(at), \quad t \in \mathbb{R}.$$  

(ii) Let $\xi$ and $\eta$ be independent random variables. Then we have

$$c_{\xi+\eta}(t) = c_{\xi}(t)c_{\eta}(t), \quad t \in \mathbb{R}.$$  

We consider some examples.

**Example 2.10** (Poisson approximation to Binomial). Suppose that $\xi_n$ has a Binomial$(n, p_n)$ distribution such that as $n \to \infty$, $np_n \to \lambda$. It can be shown that the mgf of $\xi_n$ is given by

$$m_n(t) = (p_n e^t + 1 - p_n)^n,$$

which converges to $\exp\{\lambda(e^t - 1)\}$ as $n \to \infty$ for any $t \in \mathbb{R}$ (Exercise 2.24). On the other hand, $\exp\{\lambda(e^t - 1)\}$ is the mgf of $\xi \sim \text{Poisson}(\lambda)$. Therefore, by
Theorem 2.10, we have $\xi_n \xrightarrow{d} \xi$ as $n \to \infty$. This justifies an approximation that is often taught in elementary statistics courses; that is, the Binomial$(n, p)$ distribution can be approximated by the Poisson($\lambda$) distribution, provided that $n$ is large, $p$ is small, and $np$ is approximately equal to $\lambda$.

Example 2.11. The classical CLT may be interpreted as, under regularity conditions, the sample mean of i.i.d. observations, $X_1, \ldots, X_n$, is asymptotically normal. This sometimes leads to the impression that as $n \to \infty$ (and with a suitable normalization), the limiting distribution of

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

is always normal. However, this is not true. To see a counterexample, suppose that $X_1, \ldots, X_n$ are i.i.d. with the pdf

$$f(x) = \frac{1 - \cos(x)}{\pi x^2}, \quad -\infty < x < \infty.$$

Note that the mgf of $X_i$ does not exist for any $t \neq 0$. However, the cf of $X_i$ is given by $\max(1 - |t|, 0)$, $t \in \mathbb{R}$ (Exercise 2.25). Furthermore, by Lemma 2.3 it can be shown that the cf of $\bar{X}$ is given by

$$\left\{ \max \left( 1 - \frac{|t|}{n}, 0 \right) \right\}^n,$$

which converges to $e^{-|t|}$ as $n \to \infty$ (Exercise 2.25). However, the latter is the cf of the Cauchy$(0, 1)$ distribution. Therefore, in this case, the sample mean is asymptotically Cauchy instead of asymptotically normal. The violation of the CLT is due to the failure of the regularity conditions—namely, that $X_i$ has finite expectation (and variance; see Section 6.4 for details).

In many cases, convergence in distribution of a sequence can be derived from the convergence in distribution of another sequence. We conclude this section with some useful results of this type.

**Theorem 2.12** (Continuous mapping theorem). Suppose that $\xi_n \xrightarrow{d} \xi$ as $n \to \infty$ and that $g$ is a continuous function. Then $g(\xi_n) \xrightarrow{d} g(\xi)$ as $n \to \infty$.

The proof is omitted (e.g., Billingsley 1995, §5). Alternatively, Theorem 2.12 can be derived from Theorem 2.18 given in Section 2.7 (Exercise 2.27).

**Theorem 2.13** (Slutsky’s theorem). Suppose that $\xi_n \xrightarrow{d} \xi$ and $\eta_n \xrightarrow{P} c$, as $n \to \infty$, where $c$ is a constant. Then (i) $\xi_n + \eta_n \xrightarrow{d} \xi + c$, and (ii) $\xi_n \eta_n \xrightarrow{d} c \xi$ as $n \to \infty$.

The proof is left as an exercises (Exercise 2.26).
The next result involves an extension of convergence in distribution to the multivariate case. Let \( \xi = (\xi_1, \ldots, \xi_k) \) be a random vector. The cdf of \( \xi \) is defined as
\[
F(x_1, \ldots, x_k) = P(\xi_1 \leq x_1, \ldots, \xi_k \leq x_k), \quad x_1, \ldots, x_k \in \mathbb{R}.
\]
A sequence of random vectors \( \xi_n, n = 1, 2, \ldots \), converges in distribution to a random vector \( \xi \), denoted by \( \xi_n \xrightarrow{d} \xi \), if the cdf of \( \xi_n \) converges to the cdf of \( \xi \), denoted by \( F \), at every continuity point of \( F \).

**Theorem 2.14.** Let \( \xi_n, n = 1, 2, \ldots \), be a sequence of \( d \)-dimensional random vectors. Then \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \) if and only if \( a'\xi_n \xrightarrow{d} a'\xi \) as \( n \to \infty \) for every \( a \in \mathbb{R}^d \).

### 2.5 \( L^p \) convergence and related topics

Let \( p \) be a positive number. A sequence of random variables \( \xi_n, n = 1, 2, \ldots \) converges in \( L^p \), to a random variable \( \xi \), denoted by \( \xi_n \xrightarrow{L^p} \xi \), if \( E(|\xi_n - \xi|^p) \to 0 \) as \( n \to \infty \). \( L^p \) convergence (for any \( p > 0 \)) implies convergence in probability, as the following theorem states, which can be proved by applying Chebyshev’s inequality (Exercise 2.30).

**Theorem 2.15.** \( \xi_n \xrightarrow{L^p} \xi \) implies \( \xi_n \xrightarrow{P} \xi \).

The converse, however, is not true, as the following example shows.

**Example 2.12.** Let \( X \) be a random variable that has the following pdf with respect to the Lebesgue measure
\[
f(x) = \frac{\log a}{x(\log x)^2}, \quad x \geq a,
\]
where \( a \) is a constant such that \( a > 1 \). Let \( \xi_n = X/n, n = 1, 2, \ldots \). Then we have \( \xi_n \xrightarrow{P} 0 \), as \( n \to \infty \). In fact, for any \( \epsilon > 0 \), we have
\[
P(|\xi_n| > \epsilon) = P(X > n\epsilon)
= \int_{n\epsilon}^{\infty} \frac{\log a}{x(\log x)^2} \, dx
= \frac{\log a}{\log(n\epsilon)} \xrightarrow{n \to \infty} 0
\]
as \( n \to \infty \). On the other hand, for any \( p > 0 \), we have
$$E(|\xi|^p) = E\left(\frac{X}{n}\right)^p = \frac{\log a}{n^p} \int_a^\infty \frac{dx}{x^{1-p}(\log x)^2} = \infty;$$
so it is not true that $\xi_n \overset{L^p}{\to} 0$ as $n \to \infty$.

Note that in the above example the sequence $\xi_n$ converges in probability; yet it does not converge in $L^p$ for any $p > 0$. However, the following theorem states that, under an additional assumption, convergence in probability indeed implies $L^p$ convergence.

**Theorem 2.16** (Dominated convergence theorem). Suppose that $\xi_n \overset{P}{\to} \xi$ as $n \to \infty$, and there is a nonnegative random variable $\eta$ such that $E(\eta^p) < \infty$, and $|\xi_n| \leq \eta$ for all $n$. Then $\xi_n \overset{L^p}{\to} \xi$ as $n \to \infty$.

The proof is based on the following lemma whose proof is omitted (e.g., Chow and Teicher 1988, §4.2).

**Lemma 2.4** (Fatou’s lemma). Let $\eta_n$, $n = 1, 2, \ldots$, be a sequence of random variables such that $\eta_n \geq 0$, a.s. Then

$$E(\lim inf \eta_n) \leq \lim inf E(\eta_n).$$

**Proof of Theorem 2.16.** First, we consider a special case so that $\xi_n \overset{a.s.}{\to} \xi$. Then, $|\xi| = \lim_{n \to \infty} |\xi_n| \leq \eta$, a.s. Consider $\eta_n = (2\eta)^p - |\xi_n - \xi|^p$. Since $|\xi_n - \xi| \leq |\xi_n| + |\xi| \leq 2\eta$, a.s., we have $\eta_n \geq 0$, a.s. Thus, by Lemma 2.4 and the results of §1.5.1.5, we have

$$(2\eta)^p = E(\lim inf \eta_n) \leq \lim inf E(\eta_n) = \lim inf \{(2\eta)^p - E(|\xi_n - \xi|^p)\} \leq (2\eta)^p - \lim sup E(|\xi_n - \xi|^p),$$

which implies $\lim sup E(|\xi_n - \xi|^p) \leq 0$; hence, $E(|\xi_n - \xi|^p) \to 0$ as $n \to \infty$.

Now, we drop the assumption that $\xi_n \overset{a.s.}{\to} \xi$. We use the argument of subsequences (see §1.5.1.6). It suffices to show that for any subsequence $\xi_{n_k}$, $k = 1, 2, \ldots$, there is a further subsequence $\xi_{n_{kl}}$, $l = 1, 2, \ldots$, such that

$$E\left(|\xi_{n_{kl}} - \xi|^p\right) \to 0 \quad (2.12)$$
as $l \to \infty$. Since $\xi_n \overset{P}{\to} \xi$, so does the subsequence $\xi_{n_k}$. Then, according to a result given later in Section 2.7 (see §2.7.2), there is a further subsequence
\[ n_{k_l} \text{ such that } \xi_{n_{k_l}} \xrightarrow{\text{a.s.}} \xi \text{ as } l \to \infty. \] Result (2.12) then follows from the proof given above assuming a.s. convergence. This completes the proof. Q.E.D.

The dominated convergence theorem is a useful result that is often used to establish \( L^p \) convergence given convergence in probability or a.s. convergence. We consider some examples.

**Example 2.13.** Let \( X_1, \ldots, X_n \) be i.i.d. Bernoulli\((p)\) observations. The sample proportion (or binomial proportion)
\[
\hat{p} = \frac{X_1 + \cdots + X_n}{n}
\]
converges in probability to \( p \) (it also converges a.s. according to the bounded SLLN; see Example 2.7). Since \( |X_i| \leq 1 \), by Theorem 2.16, \( \hat{p} \) converges to \( p \) in \( L^p \) for any \( p > 0 \).

**Example 2.14.** In Example 1.4 we showed that if \( X_1, \ldots, X_n \) are i.i.d. observations from the Uniform\([0, \theta]\) distribution, then the MLE of \( \theta \), \( \hat{\theta} = X_{(n)} \), is consistent; that is \( \hat{\theta} \xrightarrow{P} \theta \) as \( n \to \infty \). Because \( 0 \leq \hat{\theta} \leq \theta \), Theorem 2.16 implies that \( \hat{\theta} \) converges in \( L^p \) to \( \theta \) for any \( p > 0 \).

Another concept that is closely related to \( L^p \) convergence is called uniform integrability. The sequence \( \xi_n, n = 1, 2, \ldots, \) is uniformly integrable in \( L^p \) if
\[
\lim_{a \to \infty} \sup_{n \geq 1} E\{|\xi_n|^p 1_{(\xi_n > a)}\} = 0. \tag{2.13}
\]

**Theorem 2.17.** Suppose that \( E(|\xi_n|^p) < \infty, n = 1, 2, \ldots, \) and \( \xi_n \xrightarrow{P} \xi \) as \( n \to \infty \). Then the following are equivalent:
(i) \( \xi_n, n = 1, 2, \ldots, \) is uniformly integrable in \( L^p \);
(ii) \( \xi_n \xrightarrow{L^p} \xi \) as \( n \to \infty \) with \( E(|\xi|^p) < \infty \);
(iii) \( E(|\xi_n|^p) \to E(|\xi|^p), \) as \( n \to \infty \).

**Proof.** (i) \( \Rightarrow \) (ii): First, assume that \( \xi_n \xrightarrow{\text{a.s.}} \xi \). Then, for any \( a > 0 \), the following equality holds almost surely:
\[
|\xi|^p 1_{(|\xi| > a)} = \left\{ \lim_{n \to \infty} |\xi_n|^p 1_{(|\xi_n| > a)} \right\} 1_{(|\xi| > a)}.
\]
To see this, note that if \( |\xi| \leq a \), both sides of the equation are zero; and if \( |\xi| > a \), then \( \xi_n \to \xi \) implies that \( |\xi_n| > a \) for large \( n \); hence, \( |\xi_n|^p 1_{(|\xi_n| > a)} = |\xi|^p \to |\xi|^p \), which is the left side. Thus, by Fatou’s lemma, we have
\[
E\{|\xi|^p 1_{(|\xi| > a)}\} \leq E\left\{ \lim_{n \to \infty} |\xi_n|^p 1_{(|\xi_n| > a)} \right\}
\leq \lim inf E\{|\xi_n|^p 1_{(|\xi_n| > a)}\}
\leq \sup_{n \geq 1} E\{|\xi_n|^p 1_{(|\xi_n| > a)}\}. \tag{2.14}
\]
Furthermore, we have the argument of subsequences (Exercise 2.33). Combining (2.18) and (2.19), we have convergence theorem, we have 

\[ E(|\xi_n - \xi|^p) = E(|\xi_n - \xi|^p 1_{(|\xi_n| < a)} + |\xi_n - \xi|^p 1_{(|\xi_n| > a)} \right) \leq \left( |a|^p + \epsilon \right) < \infty. \]

Furthermore, we have

\[
|\xi_n - \xi|^p = |\xi_n - \xi|^p 1_{(|\xi_n| \leq a)} + |\xi_n - \xi|^p 1_{(|\xi_n| > a)} \leq |a|^p + \epsilon < \infty. \tag{2.15}
\]

If \( |\xi| \leq a < |\xi_n| \), then \( |\xi_n - \xi| \leq |\xi_n| + |\xi| < 2|\xi_n| \); hence, the second term on the right side of (2.15) is bounded by \( 2^p|\xi_n|^p 1_{(|\xi_n| > a)} \). On the other hand, by the inequality

\[ |u - v|^p \leq 2^p(|u|^p + |v|^p), \quad u, v \in R \quad \tag{2.16} \]

(Exercise 2.32), the third term on the right side of (2.15) is bounded by \( 2^p\{|\xi_n|^p 1_{(|\xi_n| > a)} + |\xi|^p 1_{(|\xi| > a)}\} \). Therefore, by (2.15), we have

\[
E(|\xi_n - \xi|^p) \leq E\{|\xi_n - \xi|^p 1_{(|\xi_n| \leq a)}\} + 2^p E\{|\xi|^p 1_{(|\xi| > a)}\} \leq E\{|\xi_n - \xi|^p 1_{(|\xi_n| \leq a)}\} + 3 \cdot 2^p \epsilon. \tag{2.17}
\]

Finally, we \( |\xi_n - \xi|^p 1_{(|\xi_n| \leq a)} \xrightarrow{a.s.} 0 \) and \( |\xi_n - \xi|^p 1_{(|\xi_n| > a)} \leq 2^p(a^p + |\xi|^p) \) by (2.16), and \( E(|\xi|^p) < \infty \) as is proved above. Thus, by the dominated convergence theorem, we have \( E\{|\xi_n - \xi|^p 1_{(|\xi_n| \leq a)}\} \rightarrow 0 \) as \( n \rightarrow \infty \). It follows, by (2.17) and the results of §1.5.1.5, that

\[
\limsup E(|\xi_n - \xi|^p) \leq 3 \cdot 2^p \epsilon.
\]

Since \( \epsilon \) is arbitrary, we have \( E(|\xi_n - \xi|^p) \rightarrow 0 \) as \( n \rightarrow \infty \). 

We now drop the assumption that \( \xi_n \xrightarrow{a.s.} \xi \). The result then follows by the argument of subsequences (Exercise 2.33).

(ii) \( \Rightarrow \) (iii): For any \( a > 0 \), we have

\[
|\xi_n|^p - |\xi|^p = (|\xi_n|^p - |\xi|^p) 1_{(|\xi_n| \leq a)} + (|\xi_n|^p - |\xi|^p) 1_{(|\xi_n| > a)} = \eta_n + \zeta_n. \tag{2.18}
\]

By (2.16), we have

\[
|\zeta_n| \leq |\xi_n|^p 1_{(|\xi_n| > a)} + |\xi|^p 1_{(|\xi_n| > a)} \leq 2^p(|\xi|^p + |\xi_n - \xi|^p) 1_{(|\xi_n| > a)} + |\xi|^p 1_{(|\xi_n| > a)} \leq (2^p + 1)|\xi|^p 1_{(|\xi| > a)} + 2^p|\xi_n - \xi|^p. \tag{2.19}
\]

Combining (2.18) and (2.19), we have

\[
E(|\xi_n|^p - |\xi|^p) \leq E(\eta_n) + (2^p + 1)E\{|\xi|^p 1_{(|\xi| > a)}\} + 2^pE(|\xi_n - \xi|^p) = I_1 + I_2 + I_3
\]
By Theorem 2.15, we have $\eta_n \xrightarrow{P} 0$; hence, by Theorem 2.16, we have $I_1 \to 0$ as $n \to \infty$. Also (ii) implies $I_3 \to 0$, as $n \to \infty$. Thus, we have (see §1.5.1.5)

$$\limsup E(||\xi_n|^p - |\xi|^p|) \leq (2^p + 1)E\{|\xi|^p 1(|\xi| > a)\}.$$ 

Note that $a$ is arbitrary and, by Theorem 2.16, it can be shown that

$$E\{|\xi|^p 1(|\xi| > a)\} \to 0 \quad \text{as} \quad a \to \infty.$$ 

(2.20)

This implies $E(||\xi_n|^p - |\xi|^p|) \to 0$, which implies $E(|\xi_n|^p) \to E(|\xi|^p)$ as $n \to \infty$ (Exercise 2.34).

(iii) $\Rightarrow$ (i): For any $a > 0$, we have

$$E\{|\xi_n|^p 1(|\xi_n| > a)\} = E(|\xi_n|^p) - E\{|\xi_n|^p 1(|\xi_n| \leq a)\} \leq E(|\xi_n|^p) - E\{|\xi_n|^p 1(|\xi| \leq a, |\xi| < a)\} = I_1 - I_2.$$

(iii) implies $I_1 \to E(|\xi|^p)$. Furthermore, let $\eta_n = |\xi_n|^p 1(|\xi_n| \leq a, |\xi| < a)$. It can be shown (Exercise 2.35) that $\eta_n \xrightarrow{P} \eta = |\xi|^p 1(|\xi| < a)$ as $n \to \infty$. In addition, we have $0 \leq \eta_n \leq a^p$. Thus, by Theorem 2.16, we have $\eta_n \xrightarrow{L^p} \eta$, which implies $I_2 = E(\eta_n) \to E(\eta)$. We now use the arguments of §1.5.1.5 to conclude that

$$\limsup E\{|\xi_n|^p 1(|\xi_n| > a)\} \leq E(|\xi|^p) - E\{|\xi|^p 1(|\xi| < a)\} = E\{|\xi|^p 1(|\xi| \geq a)\}.$$ 

(2.21)

For any $\epsilon > 0$, by (2.21) and the definition of lim sup, there is $N \geq 1$ such that $E\{|\xi_n|^p 1(|\xi_n| > a)\} \leq E\{|\xi|^p 1(|\xi| \geq a)\} + \epsilon$ if $n \geq N$. It follows that

$$\sup_{n \geq 1} E\{|\xi_n|^p 1(|\xi_n| > a)\} \leq \left[\max_{1 \leq n \leq N - 1} E\{|\xi_n|^p 1(|\xi_n| > a)\}\right] \vee \left[E\{|\xi|^p 1(|\xi| \geq a)\} + \epsilon \right].$$

Furthermore, by the dominated convergence theorem it can be shown that $E\{|\xi_n|^p 1(|\xi_n| > a)\} \to 0$, $1 \leq n \leq N - 1$, and $E\{|\xi|^p 1(|\xi| \geq a)\} \to 0$ as $a \to \infty$ (see Exercise 2.34). Therefore, we have

$$\limsup_{n \geq 1} \sup_{n \geq 1} E\{|\xi_n|^p 1(|\xi_n| > a)\} \leq \epsilon,$$

where the lim sup is with respect to $a$. Since $\epsilon$ is arbitrary, we conclude that $\sup_{n \geq 1} E\{|\xi_n|^p 1(|\xi_n| > a)\} \to 0$ as $a \to \infty$. This completes the proof. Q.E.D.

**Example 2.15.** Suppose that $\xi_n \xrightarrow{P} \xi$ as $n \to \infty$, and that $E(|\xi_n|^q)$, $n \geq 1$, is bounded for some $q > 0$. Then $\xi_n \xrightarrow{L^p} \xi$ as $n \to \infty$ for any $0 < p < q$. To see this, note that for any $a > 0$, $|\xi_n| > a$ implies $|\xi_n|^p = a^{p-q}$. Thus,

$$E\{|\xi_n|^p 1(|\xi_n| > a)\} \leq a^{p-q}E(|\xi|^q) \leq Ba^{p-q},$$
where $B = \sup_{n \geq 1} \mathbb{E}(|\xi_n|^p) < \infty$. Because $p - q < 0$, we have
\[
\sup_{n \geq 1} \mathbb{E}\{\xi_n^p 1(\xi_n > a)\} \to 0
\]
as $a \to \infty$. In other words, $\xi_n, n = 1, 2, \ldots$, is uniformly integrable. The result then follows by Theorem 2.17.

**Example 2.16.** Let $X$ be a random variable that has a pdf $f(x)$ with respect to a $\sigma$-finite measure $\mu$ (see Appendix A.2). Suppose that $f_n(x), n = 1, 2, \ldots$, is a sequence of pdf’s with respect to $\mu$ such that $f_n(x) \to f(x), x \in R$, as $n \to \infty$. Consider the sequence of random variables
\[
\xi_n = \frac{f_n(X)}{f(X)} \quad (2.22)
\]
n = 1, 2, \ldots. Then we have $\xi_n \overset{L^1}{\to} 1$ as $n \to \infty$. To see this, note that $f_n(x) \to f(x), x \in R$ implies $\xi_n \overset{a.s.}{\to} 1$. This is because $f_n(x)/f(x) \to 1$ as long as $f(x) > 0$; hence, $\mathbb{P}(\xi_n \to 1) \geq \mathbb{P}\{f(X) > 0\} = 1 - \mathbb{P}\{f(X) = 0\}$ and
\[
\mathbb{P}\{f(X) = 0\} = \int_{f(x)=0} f(x) \, d\mu = 0.
\]
It follows by Theorem 2.7 that $\xi_n \overset{p}{\to} 1$. On the other hand, we have
\[
\mathbb{E}(|\xi_n|) = \mathbb{E}\left\{\frac{f_n(X)}{f(X)}\right\} = \int \frac{f_n(x)}{f(x)} f(x) \, d\mu = \int f_n(x) \, d\mu = 1.
\]
Thus, by Theorem 2.17, we have $\xi_n \overset{L^1}{\to} 1$ as $n \to \infty$.

When $X$ is a vector of observations, $(2.22)$ corresponds to a likelihood ratio, which may be thought as the probability of observing $X$ under $f_n$ divided by that under $f$. Thus, the above example indicates that if $f_n$ converges to $f$ pointwisely, then the likelihood ratio converges to $1$ in $L^1$, provided that $f(x)$ is the true pdf of $X$. To see a specific example, suppose that $X$ has a standard normal distribution; that is, $X \sim f(x)$, where
\[
f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.
\]
Let $f_n(x)$ be the pdf of the $t$-distribution with $n$ degrees of freedom; that is,
\[ f_n(x) = \frac{\Gamma\{(n + 1)/2\}}{\sqrt{n\pi} \Gamma(n/2)} \left( 1 + \frac{x^2}{n} \right)^{-\frac{(n+1)/2}{2}}, \quad -\infty < x < \infty. \]

Then, by Exercise 1.4, we have \( f_n(x) \rightarrow f(x), \ x \in \mathbb{R}, \) as \( n \rightarrow \infty. \) It follows that \( f_n(X)/f(X) \overset{L^1}{\rightarrow} 1 \) as \( n \rightarrow \infty. \) It should be pointed out that the \( L^1 \) convergence may not hold if \( f(x) \) is not the true distribution of \( X, \) even if \( f_n(x) \rightarrow f(x) \) for every \( x. \) For example, suppose that in the Example 2.16 involving the \( t \)-distribution, the distribution of \( X \) is \( N(0,2) \) instead of \( N(0,1); \) then, clearly, we still have \( f_n(x) \rightarrow f(x), \ x \in \mathbb{R} \) \( \{f(x) \) has not changed; only that \( X \sim f(x) \) no longer holds\}. However, it is not true that \( f_n(X)/f(X) \overset{L^1}{\rightarrow} 1. \) This is because, otherwise, by the inequality

\[ \frac{f_n(X)}{f(X)} \leq 1 + \left| \frac{f_n(X)}{f(X)} - 1 \right|, \]

we would have

\[ \mathbb{E} \left\{ \frac{f_n(X)}{f(X)} \right\} \leq 1 + \mathbb{E} \left( \left| \frac{f_n(X)}{f(X)} - 1 \right| \right) \]

\[ \leq 2 \]

for large \( n. \) However,

\[ \mathbb{E} \left\{ \frac{f_n(X)}{f(X)} \right\} = \int \frac{f_n(x)}{f(x)} \frac{1}{\sqrt{4\pi}} e^{-x^2/4} dx \]

\[ = \frac{1}{\sqrt{2}} \int \frac{\Gamma\{(n + 1)/2\}}{\sqrt{n\pi} \Gamma(n/2)} \left( 1 + \frac{x^2}{n} \right)^{-\frac{(n+1)/2}{2}} e^{x^2/4} dx \]

\[ = \infty. \]

We conclude this section by revisiting the example that began the section.

**Example 2.1 (continued).** It is clear now that CLT means convergence in distribution—that is, \( \xi_n = \sqrt{n}(\bar{X} - \mu) \overset{d}{\rightarrow} \xi \sim N(0, \sigma^2) \)—but this does not necessarily imply \( \text{var}(\sqrt{n}\bar{X}) = \mathbb{E}(\xi_n^2) \rightarrow \mathbb{E}(\xi^2) = \sigma^2 \) (see an extension of parts of Theorem 2.17 in Section 2.7, where the convergence in probability condition is weakened to convergence in distribution). In fact, the CLT even holds in some situations where the variance of the \( X_i \)’s do not exist (see Chapter 6).

### 2.6 Case study: \( \chi^2 \)-test

One of the celebrated results in classical statistics is Pearson’s \( \chi^2 \) goodness-of-fit test, or simply \( \chi^2 \)-test (Pearson 1900). The test statistic is given by

\[ \chi^2 = \sum_{k=1}^{M} \frac{(O_k - E_k)^2}{E_k}, \quad (2.23) \]
where $M$ is the number of cells into which $n$ observations are grouped, $O_k$ and $E_k$ are the observed and expected frequencies of the $k$th cell, $1 \leq k \leq M$, respectively. The expected frequency of the $k$th cell is given by $E_k = np_k$, where $p_k$ is the known cell probability of the $k$th cell evaluated under the assumed model. The asymptotic theory associated with this test is simple: Under the null hypothesis of the assumed model, $\chi^2 \overset{d}{\to} \chi^2_{M-1}$ as $n \to \infty$.

One good feature of Pearson’s $\chi^2$-test is that it can be used to test an arbitrary probability distribution, provided that the cell probabilities are completely known. However, the latter actually is a serious constraint, because in practice the cell probabilities often depend on certain unknown parameters of the probability distribution specified by the null hypothesis. For example, under the normal null hypothesis, the cell probabilities depend on the mean and variance of the normal distribution, which may be unknown. In such a case, intuitively one would replace the unknown parameters by their estimators and thus obtain the estimated $\hat{E}_k$, say $\hat{E}_k$, $1 \leq k \leq M$. The test statistic (2.23) then becomes

$$\hat{\chi}^2 = \sum_{k=1}^{M} \left( \frac{O_k - \hat{E}_k}{\hat{E}_k} \right)^2.$$

(2.24)

However, this test statistic may no longer have an asymptotic $\chi^2$-distribution.

In a simple problem of assessing the goodness-of-fit to a Poisson or Multinomial distribution, it is known that the asymptotic null-distribution of (2.24) is $\chi^2_{M-p-1}$, where $p$ is the number of parameters estimated by the maximum likelihood method. This is the famous “subtract one degree of freedom for each parameter estimated” rule taught in many elementary statistics books (e.g., Rice 1995, pp. 242). However, the rule may not be generalizable to other probability distributions. For example, this rule does not even apply to testing normality with unknown mean and variance, as mentioned above. Note that here we are talking about MLE based on the original data, not the MLE based on cell frequencies. It is known that the rule applies in general to MLE based on cell frequencies. However, the latter are less efficient than the MLE based on the original data except for special cases where the two are the same, such as the above Poisson and Multinomial cases.

R. A. Fisher was the first to note that the asymptotic null-distribution of (2.24) is not necessarily $\chi^2$ (Fisher 1922a). He showed that if the unknown parameters are estimated by the so-called minimum chi-square method, the asymptotic null-distribution of (2.24) is still $\chi^2_{M-p-1}$, but this conclusion may be false if other methods of estimation (including the ML) are used. Note that there is no contradiction of Fisher’s result with the above results related to Poisson and Multinomial distributions, because the minimum chi-square estimators and the MLE are asymptotically equivalent when both are based on cell frequencies. A more thorough result was obtained by Chernoff and Lehmann (1954), who showed that when the MLE based on the original observations are used, the asymptotic null-distribution of (2.24) is not necessarily
2.6 Case study: $\chi^2$-test

$\chi^2$, but instead a “weighted” $\chi^2$, where the weights are eigenvalues of certain nonnegative definite matrix. Note that the problem is closely related to the first example given in the Preface of this book. See Moore (1978) for a nice historical review of the $\chi^2$-test.

There are two components in Pearson’s $\chi^2$-test: the (observed) cell frequencies, $O_k$, $1 \leq k \leq M$, and the cell probabilities, $p_k$, $1 \leq k \leq M$. Although considerable attention has been given to address the issue associated with the $\chi^2$-test with estimated cell probabilities, there are situations in practice where the cell frequencies also need to be estimated. The following is an example.

**Example 2.17** (Nested-error regression). Consider a situation of clustered observations. Let $Y_{ij}$ denote the $j$th observation in the $i$th cluster. Suppose that $Y_{ij}$ satisfies the following nested-error regression model:

$$Y_{ij} = x_{ij}' \beta + u_i + e_{ij},$$

$i = 1, \ldots, n$, $j = 1, \ldots, b$, where $x_{ij}$ is a known vector of covariates, $\beta$ is an unknown vector of regression coefficients, $u_i$ is a random effect, and $e_{ij}$ is an additional error term. It is assumed that the $u_i$’s are i.i.d. with distribution $F$ that has mean 0, the $e_{ij}$’s are i.i.d. with distribution $G$ that has mean 0, and the $u_i$’s and $e_{ij}$’s are independent. Here, both $F$ and $G$ are unknown.

Jiang, Lahiri, and Wu (2001) extended Pearson’s $\chi^2$-test to situations where both the cell frequencies and cell probabilities have to be estimated. In the remaining part of this section we describe their approach without giving all of the details. The details are referred to the reference above. Let $Y$ be a vector of observations whose joint distribution depends on an unknown
vector of parameters, \( \theta \). Suppose that \( X_i(\theta) = X_i(y, \theta) \) satisfy the following conditions: (i) for any fixed \( \theta \), \( X_1(\theta), \ldots, X_n(\theta) \) are independent; and (ii) if \( \theta \) is the true parameter vector, \( X_1(\theta), \ldots, X_n(\theta) \) are i.i.d.

**Example 2.17 (continued).** If we let \( \theta = \beta \) and \( X_i(\theta) = Y_i - \bar{X}_i / \beta, 1 \leq i \leq n \), then conditions (i) and (ii) are satisfied (Exercise 2.36).

Let \( C_k, 1 \leq k \leq M \) be disjoint subsets of \( R \) such that \( \bigcup_{k=1}^{M} C_k \) covers the range of \( X_i(\theta), 1 \leq i \leq n \). Define \( p_{i,k}(\theta, \tilde{\theta}) = P_{\theta} \{ X_i(\tilde{\theta}) \in C_k \}, 1 \leq k \leq M \), and \( p_i(\tilde{\theta}) = [p_{i,k}(\theta, \tilde{\theta})]_{1 \leq k \leq M} \). Here, \( P_{\theta} \) denotes the probability given that \( \theta \) is the true parameter vector. Note that under assumption (ii), \( p_i(\tilde{\theta}) \) does not depend on \( i \) (why?). Therefore, it will be denoted by \( p(\theta) = [p_k(\theta)]_{1 \leq k \leq M} \).

If \( \theta \) were known, one would have observed \( X_i(\theta) \) and hence compute the \( \chi^2 \) statistic (2.24); that is,

\[
\hat{\chi}^2_0 = \sum_{k=1}^{M} \frac{(O_k(\theta) - np_k(\theta))^2}{np_k(\theta)},
\]

(2.26)

where \( O_k(\theta) = \sum_{i=1}^{n} 1_{\{X_i(\theta) \in C_k\}} \). Here, \( p_k(\theta) \) is computed under the null hypothesis. However, \( O_k(\theta) \) is not observable, because \( \theta \) is unknown. Instead, we compute an estimated cell frequency, \( \hat{O}_k(\hat{\theta}) = \sum_{i=1}^{n} 1_{\{X_i(\hat{\theta}) \in C_k\}} \), where \( \hat{\theta} \) is an estimator of \( \theta \). If we replace \( O_k(\theta) \) by \( \hat{O}_k(\hat{\theta}) \) and \( p_k(\theta) \) by \( p_k(\hat{\theta}) \) in (2.26), we come up with the following \( \chi^2 \) statistic:

\[
\hat{\chi}^2_e = \sum_{k=1}^{M} \frac{(\hat{O}_k(\hat{\theta}) - np_k(\hat{\theta}))^2}{np_k(\hat{\theta})},
\]

(2.27)

Here, the subscript \( e \) represents “estimated” (frequencies).

Our goal is to obtain the asymptotic distribution of \( \hat{\chi}^2_e \). In order to do so, we need some regularity conditions, including assumptions about \( \hat{\theta} \). We assume that \( p_i(\theta, \tilde{\theta}) \) is twice times continuously differentiable with respect to \( \theta \) and \( \tilde{\theta} \). Let \( \theta \) denotes the true parameter vector. We assume that \( p_k(\theta) > 0, 1 \leq k \leq M \), and there is \( \delta > 0 \) such that the following are bounded:

\[
\sup_{|\tilde{\theta} - \theta| < \delta} \left\| \frac{\partial}{\partial \theta^\prime} p_i(\theta, \tilde{\theta}) \right\|,
\]

\[
\sup_{|\tilde{\theta} - \theta| < \delta} \left\| \frac{\partial^2}{\partial \theta \partial \theta^\prime} p_i(\theta, \tilde{\theta}) \right\|,
\]

\[
\sup_{|\theta_1 - \theta| < \delta, |\theta_2 - \theta| < \delta} \left\| \frac{\partial^2}{\partial \theta_1 \partial \theta_2} p_i,k(\theta_1, \theta_2) \right\|,
\]

\[
1 \leq k \leq M, 1 \leq i \leq n \text{ (see Appendix A.1 for notation of matrix norms and differentiation).}
\]

Furthermore, we assume that for fixed \( \tilde{\theta}, X_i(\tilde{\theta}), 1 \leq i \leq n, \) are independent of \( \tilde{\theta} \), and \( \tilde{\theta} \) satisfies
\[
\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\{0, A(\theta)\},
\]
where the covariance matrix \(A(\theta)\) may be singular. Then the asymptotic distribution of \(\chi^2_e\) is the same as the distribution of
\[
\sum_{j=1}^r (1 + \lambda_j)Z_j^2 + \sum_{j=r+1}^{M-1} Z_j^2,
\]
where \(r = \text{rank}\{B(\theta)\}\) with
\[
B(\theta) = \text{diag}\{p(\theta)\}^{-1/2}Q(\theta)A(\theta)Q(\theta)'\text{diag}\{p(\theta)\}^{-1/2},
\]
\[
Q(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left. \frac{\partial}{\partial \theta'} p_i(\theta, \hat{\theta}) \right|_{\hat{\theta} = \theta}
\]
(see Appendix A.4 for notation), \(\lambda_j, 1 \leq j \leq r\) are the positive eigenvalues of \(B(\theta)\), and \(Z_j, 1 \leq j \leq M - 1\) are independent \(N(0, 1)\) random variables.

Note that in spite of the fact that \(p_i(\theta, \theta) = p(\theta)\), \((\partial/\partial \theta')p_i(\theta, \hat{\theta})\big|_{\hat{\theta} = \theta}\) is not necessarily equal to \((\partial/\partial \theta')p(\theta)\) (Exercise 2.37). Therefore, the right side of (2.31) is not necessarily equal to \((\partial/\partial \theta')p(\theta)\).

Comparing the above result with the well-known results about the \(\chi^2\)-test (e.g., Chernoff and Lehmann 1954), we observe the following:

(i) If no parameter is estimated, the asymptotic distribution of \(\hat{\chi}^2_0\), defined by (2.26), is the same as that of \(\sum_{j=1}^{M-1} Z_j^2\).

(ii) If the parameters are estimated by the MLE based on the cell frequencies, the asymptotic distribution of the resulting \(\chi^2\) statistic, say, \(\hat{\chi}^2_2\), is the same as that of \(\sum_{j=1}^{M-s-1} Z_j^2\), where \(s\) is the number of (independent) parameters estimated.

(iii) If the parameters are estimated by the MLE based on the original data, the asymptotic distribution of the resulting \(\chi^2\) statistic, say, \(\hat{\chi}^2_1\), is the same as that of \(\sum_{j=1}^{M-s-1} Z_j^2 + \sum_{j=M-s}^{M-1} \mu_j Z_j^2\), where \(0 \leq \mu_j \leq 1, M - s \leq j \leq M - 1\).

It is interesting to note that, stochastically, we have
\[
\sum_{j=1}^{M-s-1} Z_j^2 \leq \sum_{j=1}^{M-s-1} Z_j^2 + \sum_{j=M-s}^{M-1} \mu_j Z_j^2 \\
\leq \sum_{j=1}^{M-1} Z_j^2 \leq \sum_{j=1}^{M-1} (1 + \lambda_j)Z_j^2 + \sum_{j=r+1}^{M-1} Z_j^2.
\]

The interpretation is the following. In \(\hat{\chi}^2_1\) and \(\hat{\chi}^2_2\), \(\hat{\theta}\) is computed from the same data, whereas in \(\chi^2_e\), \(\hat{\theta}\) is obtained from an independent source. When using the same data to compute the cell frequencies and estimate \(\theta\), the overall variation tends to reduce. To see this, consider a simple example in which \(X_1, \ldots, X_n\) are i.i.d. \(\sim \text{Bernoulli}(p)\), where \(p\) is unknown. The observed frequency for \(X_1 = 1\) is \(O_1 = \sum_{i=1}^n X_i\); the expected frequency is \(E_1 = np\), so \((O_1 - E_1)^2 = (O_1 - np)^2\).
However, if one estimates \( p \) by its MLE, \( \hat{p} = O_1/n \), one has \( \hat{E}_1 = n\hat{p} = O_1 \); therefore, \( (O_1 - \hat{E}_1)^2 = 0 \) (i.e., there is no variation). On the other hand, if \( \hat{\theta} \) is obtained from an independent source, it introduces additional variation, which is the implication of (2.32).

The assumption that \( \hat{\theta} \) is independent with \( X_i(\bar{\theta}) \), \( 1 \leq i \leq n \), for fixed \( \bar{\theta} \) may seem a bit restrictive. On the other hand, in some cases, information obtained from previous studies can be used to obtain \( \hat{\theta} \). In such a case, it may be reasonable to assume that \( \hat{\theta} \) is independent with \( X_i(\bar{\theta}) \), \( 1 \leq i \leq n \), if the latter are computed from the current data. Another situation that would satisfy the independence requirement is when \( \hat{\theta} \) is obtained by data-splitting. See Jiang, Lahiri, and Wu (2001).

We give an outline of the derivation of the asymptotic distribution of \( \hat{\chi}^2_e \). The detail can be found in Jiang, Lahiri, and Wu (2001). First, note that \( \hat{\chi}^2_e = |\xi_n|^2 \), where \( \xi_n \) is the random vector \( \text{diag}\{np(\hat{\theta})\}^{-1/2}\{O(\hat{\theta}) - np(\hat{\theta})\} \) with \( O(\theta) = [Q_k(\theta)]_{1 \leq k \leq M} \). The first step is to show

\[
\xi_n \xrightarrow{d} N(0, \Sigma(\theta) + B(\theta)) \quad (2.33)
\]
as \( n \to \infty \), where

\[
\Sigma(\theta) = I_M - \left\{ p(\theta)^{1/2} \right\} \left\{ p(\theta)^{1/2} \right\}',
\]
with \( p(\theta)^{1/2} = [p_k(\theta)^{1/2}]_{1 \leq k \leq M} \). By Theorem 2.14, (2.33) is equivalent to

\[
\lambda'\xi_n \xrightarrow{d} N[0, \lambda'\{\Sigma(\theta) + B(\theta)\}\lambda] \quad (2.34)
\]
as \( n \to \infty \) for any \( \lambda \in \mathbb{R}^M \). According to Theorem 2.11, (2.34) is, in turn, equivalent to that the cf of \( \lambda'\xi_n \) converges to the cf of the right side of (2.34). However, this is equivalent to

\[
\mathbb{E}\{\exp(i\lambda'\xi_n)\} \longrightarrow \exp\left[-\frac{1}{2}\lambda'\{\Sigma(\theta) + B(\theta)\}\lambda\right] \quad (2.35)
\]
as \( n \to \infty \) for any \( \lambda \in \mathbb{R}^M \) (Exercise 2.38). To show (2.35), we express \( \xi_n \) as

\[
\xi_n = \eta_n + \zeta_n + \gamma_n, \quad (2.36)
\]
where

\[
\eta_n = \text{diag}\{np(\hat{\theta})\}^{-1/2}\left\{O(\hat{\theta}) - \sum_{i=1}^{n} p_i(\theta, \hat{\theta})\right\},
\]
\[
\zeta_n = -\text{diag}\{p(\hat{\theta})\}^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta'} p_i(\theta, \hat{\theta})\bigg|_{\theta=\hat{\theta}}\right\} n^{1/2}(\hat{\theta} - \theta),
\]
and \( \gamma_n \) satisfies the following: There is a constant \( c \) such that for any \( \epsilon > 0 \),
\[ |\gamma_n| \leq c(n\hat{\theta} - \theta)^2 n^{-1/2} \text{ if } |\hat{\theta} - \theta| \leq \epsilon. \]

The idea for the proof of (2.35) is therefore to show that, as \( n \to \infty \), \( \eta_n \) and \( \zeta_n \) are the leading terms in (2.36) and \( \gamma_n \) is negligible. In fact, the contribution of \( \Sigma(\theta) \) in the asymptotic covariance matrix, \( \Sigma(\theta) + B(\theta) \) in (2.33), comes from \( \eta_n \) and the contribution of \( B(\theta) \) from \( \zeta_n \) (and \( \gamma_n \) has no contribution).

In other words, we have
\[
\chi'(\eta_n + \zeta_n) \lambda \xrightarrow{d} N[0, \chi'(\Sigma(\theta) + B(\theta)) \lambda]
\]
and \( \chi' \gamma_n \xrightarrow{P} 0 \) as \( n \to \infty \). Result (2.35) then follows from Slutsky’s theorem (Theorem 2.13).

Given that (2.33) holds, we apply Theorem 2.12 (note that a multivariate version of the result also stands—that is, when \( \xi_n, \xi \) are random vectors) to conclude
\[
\hat{\chi}_e^2 = |\xi_n|^2 \xrightarrow{d} |\xi|^2,
\]
where \( \xi \sim N\{0, \Sigma(\theta) + B(\theta)\} \). It remains to determine the distribution of \( |\xi^2| = \xi'\xi \). Write \( \Sigma = \Sigma(\theta) \) and \( B = B(\theta) \) and let \( P = \{p(\theta)^{1/2}\} \{p(\theta)^{1/2}\}' \). It can be shown (Exercise 2.39) that \( PB = BP = 0 \). Thus (see Appendix A.1), there is an orthogonal matrix \( T \) such that
\[
B = T \text{ diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) T',
\]
\[
P = T \text{ diag}(\rho_1, \ldots, \rho_M) T',
\]
where \( \lambda_j > 0, 1 \leq j \leq r = \text{rank}(B) \), and \( \rho_1, \ldots, \rho_M \) are the eigenvalues of \( P \). Note that the latter is a projection matrix with rank 1. Therefore, \( \rho_1, \ldots, \rho_M \) are zero except for one of them, which is 1. It follows that the distribution of \( \xi'\xi \) is the same as that of (2.29) (Exercise 2.39).

### 2.7 Summary and additional results

This section provides a summary of some of the main results in this chapter as well as some additional results. The summary focuses on the connection between different types of convergence.

1. Almost sure (a.s.) convergence implies convergence in probability, which, in turn, implies convergence in distribution.

2. If \( \xi_n \xrightarrow{P} \xi \) as \( n \to \infty \), then there is a subsequence \( n_k, k = 1, 2, \ldots \), such that \( \xi_{n_k} \xrightarrow{a.s.} \xi \) as \( k \to \infty \).

3. \( \xi_n \xrightarrow{P} \xi \) as \( n \to \infty \) if and only if for every subsequence \( n_k, k = 1, 2, \ldots \), there is a further subsequence \( n_{k_l}, l = 1, 2, \ldots \), such that
\[
\xi_{n_{k_l}} \xrightarrow{a.s.} \xi \text{ as } l \to \infty.
\]
4. If for every \( \epsilon > 0 \) we have \( \sum_{n=1}^{\infty} P(|\xi_n - \xi| \geq \epsilon) < \infty \), then \( \xi_n \xrightarrow{a.s.} \xi \) as \( n \to \infty \). Intuitively, this result states that convergence in probability at a certain rate implies a.s. convergence.

The proof of the above result follows from the following lemma, which is often useful in establishing a.s. convergence (Exercise 2.40). Let \( A_1, A_2, \ldots \) be a sequence of events. Define \( \limsup A_n = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n \).

**Lemma 2.5.** (Borel–Cantelli lemma)
(i) If \( \sum_{n=1}^{\infty} P(A_n) < \infty \), then \( P(\limsup A_n) = 0 \).
(ii) If \( A_1, A_2, \ldots \) are pairwise independent and \( \sum_{n=1}^{\infty} P(A_n) = \infty \), then \( P(\limsup A_n) = 1 \).

5. \( L^p \) convergence for any \( p > 0 \) implies convergence in probability.

6. (Dominated convergence theorem) If \( \xi_n \xrightarrow{P} \xi \) as \( n \to \infty \) and there is a random variable \( \eta \) such that \( E(\eta^p) < \infty \) and \( |\xi_n| \leq \eta, n \geq 1 \), then \( \xi_n \xrightarrow{L^p} \xi \) as \( n \to \infty \) and \( E(|\xi|^p) < \infty \).

Let \( a_n, n = 1, 2, \ldots \), be a sequence of constants. The sequence converges increasingly to \( a \), denoted by \( a_n \uparrow a \), if \( a_n \leq a_{n+1}, n \geq 1 \) and \( \lim_{n \to \infty} a_n = a \). Similarly, let \( \xi_n, n = 1, 2, \ldots \), be a sequence of random variables. The sequence converges increasingly a.s. to \( \xi \), denoted by \( \xi_n \uparrow \xi \) a.s., if \( \xi_n \leq \xi_{n+1} \) a.s., \( n \geq 1 \), and \( \lim_{n \to \infty} \xi_n = \xi \) a.s.

7. (Monotone convergence theorem) If \( \xi_n \uparrow \xi \) a.s. and \( \xi_n \geq \eta \) a.s. with \( E(|\eta|) < \infty \), then \( E(\xi_n) \uparrow E(\xi) \). The result does not imply, however, that \( E(\xi) \) is finite. So, if \( E(\xi) = \infty \), then \( E(\xi_n) \uparrow \infty \). On the other hand, we must have \( E(\xi) > -\infty \) (why?).

8. If \( \sum_{n=1}^{\infty} E(|\xi_n - \xi|^p) < \infty \) for some \( p > 0 \), then \( \xi_n \xrightarrow{a.s.} \xi \) as \( n \to \infty \). Intuitively, this means that \( L^p \) convergence at a certain rate implies a.s. convergence (Exercise 2.40).

The following theorem is useful in establishing the connection between convergence in distribution and other types of convergence.

**Theorem 2.18** (Skorokhod representation theorem). If \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \), then there are random variables \( \eta_n, n = 1, 2, \ldots \), and \( \eta \) defined on a common probability space such that \( \eta_n \) has the same distribution as \( \xi_n, n = 1, 2, \ldots \), and \( \eta \) has the same distribution as \( \xi \), and \( \eta_n \xrightarrow{a.s.} \eta \) as \( n \to \infty \).

With Skorokhod’s theorem we can extend part of Theorem 2.17 as follows.

9. If \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \), then the following are equivalent:
   (i) \( \xi_n, n = 1, 2, \ldots \), is uniformly integrable in \( L^p \).
   (ii) \( E(|\xi_n|^p) \to E(|\xi|^p) < \infty \) as \( n \to \infty \).

10. \( \xi_n \xrightarrow{d} \xi \) as \( n \to \infty \) is equivalent to \( c_n(t) \to c(t) \) as \( n \to \infty \) for every \( t \in \mathbb{R} \), where \( c_n(t) \) is the cf of \( \xi_n, n = 1, 2, \ldots \), and \( c(t) \) the cf of \( \xi \).
11. If there is $\delta > 0$ such that the mgf of $\xi_n$, $m_n(t)$, converges to $m(t)$ as $n \to \infty$ for all $t$ such that $|t| < \delta$, where $m(t)$ is the mgf of $\xi$, then $\xi_n \xrightarrow{d} \xi$ as $n \to \infty$.

12. $\xi_n \xrightarrow{d} \xi$ is equivalent to any of the following:
   (i) $\lim_{n \to \infty} \mathbb{E}\{h(\xi_n)\} = \mathbb{E}\{h(\xi)\}$ for every bounded continuous function $h$.
   (ii) $\limsup P(\xi_n \in C) \leq P(\xi \in C)$ for any closed set $C$.
   (iii) $\liminf P(\xi_n \in O) \geq P(\xi \in O)$ for any open set $O$.

13. Let $f_n(x)$ and $f(x)$ be the pdfs of $\xi_n$ and $\xi$, respectively, with respect to a $\sigma$-finite measure $\mu$ (see Appendix A.2). If $f_n(x) \to f(x)$ a.e. $\mu$ as $n \to \infty$, then $\xi_n \xrightarrow{d} \xi$ as $n \to \infty$.

14. Let $g$ be a continuous function. Then we have the following:
   (i) $\xi_n \xrightarrow{a.s.} \xi$ implies $g(\xi_n) \xrightarrow{a.s.} g(\xi)$ as $n \to \infty$;
   (ii) $\xi_n \xrightarrow{P} \xi$ implies $g(\xi_n) \xrightarrow{P} g(\xi)$ as $n \to \infty$;
   (iii) $\xi_n \xrightarrow{d} \xi$ implies $g(\xi_n) \xrightarrow{d} g(\xi)$ as $n \to \infty$.

15. (Slutsky’s theorem) If $\xi_n \xrightarrow{d} \xi$ and $\eta_n \xrightarrow{P} c$, where $c$ is a constant, then the following hold:
   (i) $\xi_n + \eta_n \xrightarrow{d} \xi + c$;
   (ii) $\eta_n \xi_n \xrightarrow{d} c\xi$;
   (iii) $\xi_n / \eta_n \xrightarrow{d} \xi / c$, if $c \neq 0$.

2.8 Exercises

2.1. Complete the definition of the sequence of random variables $\xi_n$, $n = 1, 2, \ldots$, in Example 2.1 (i.e., define $\xi_n$ for a general index $n$). Show that $\xi_n \xrightarrow{P} 0$ as $n \to \infty$; however, $\xi_n(x)$ does not converge pointwisely at any $x \in [0, 1]$.

2.2. Use Chebyshev’s inequality (see Section 5.2) to prove Theorem 2.1.

2.3. Use the $\epsilon$-$\delta$ argument to prove Theorem 2.2.

2.4. Use the $\epsilon$-$\delta$ argument to prove Theorem 2.3.

2.5. Use the $\epsilon$-$\delta$ argument to prove Theorem 2.4.

2.6. Use Theorem 2.5 and the $\epsilon$-$\delta$ argument to prove Theorem 2.6.

2.7. Let $X_1, \ldots, X_n$ be independent random variables with a common distribution $F$. Define

$$\xi_n = \frac{\max_{1 \leq i \leq n} |X_i|}{a_n}, \quad n \geq 1,$$

where $a_n$, $n = 1, 2, \ldots$, is a sequence of positive constants. Determine $a_n$ for the following cases such that $\xi_n \xrightarrow{P} 0$ as $n \to \infty$:
   (i) $F$ is the Uniform[0,1] distribution.
   (ii) $F$ is the Exponential(1) distribution.
   (iii) $F$ is the $N(0,1)$ distribution.
(iv) $F$ is the Cauchy($0, 1$) distribution.

2.8. Continue with Problem 2.7 with $a_n = n$. Show the following:

(i) If $E(|X_1|) < \infty$, then $\xi_n \xrightarrow{L^1} 0$ as $n \to \infty$.
(ii) If $E(X_1^2) < \infty$, then $\xi_n \xrightarrow{a.s.} 0$ as $n \to \infty$.

Hint: For (i), first show that for any $a > 0$,
\[
\max_{1 \leq i \leq n} |X_i| \leq a + \sum_{i=1}^{n} |X_i| 1(|X_i| > a).
\]

For (ii), use Theorem 2.8 and also note that by exchanging the order of summation and expectation, one can show for any $\epsilon > 0$,
\[
\sum_{n=1}^{\infty} nP(|X_1| > \epsilon n) < \infty.
\]

2.9. Suppose that for each $1 \leq j \leq k$, $\xi_{n,j}$, $n = 1, 2, \ldots$, is a sequence of random variables such that $\xi_{n,j} \xrightarrow{P} 0$ as $n \to \infty$. Define $\xi_n = \max_{1 \leq j \leq k} |\xi_{n,j}|$.

(i) Show that if $k$ is fixed, then $\xi_n \xrightarrow{P} 0$ as $n \to \infty$.
(ii) Give an example to show that if $k$ increases with $n$ (i.e., $k = k_n \to \infty$ as $n \to \infty$), the conclusion of (i) may not be true.

2.10. Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables. Show that $\xi_n \xrightarrow{P} 0$ as $n \to \infty$ if and only if
\[
E \left( \frac{|\xi_n|}{1 + |\xi_n|} \right) \longrightarrow 0 \quad \text{as} \quad n \to \infty.
\]

2.11. Prove Lemma 2.1 using the $\epsilon$-$\delta$ argument. Then use Lemma 2.1 to establish Theorem 2.7.

2.12. Show by similar arguments as in Example 2.7 that $I_2 \leq ce^{-\epsilon\sqrt{n}}$, where the notations refer to Example 2.7.

2.13. Verify that the infinite series $\sum_{i=1}^{\infty} e^{-\epsilon\sqrt{n}}$ converges. This result was used at the end of Example 2.7.

2.14. Suppose that $X_1, \ldots, X_n$ are i.i.d. observations with finite expectation. Show that in the following cases the sample mean $\bar{X} = (X_1 + \cdots + X_n)/n$ is a strongly consistent estimator of the population mean, $\mu = E(X_1)$—that is, $\bar{X} \xrightarrow{a.s.} \mu$ as $n \to \infty$.

(i) $X_1 \sim Binomial(m, p)$, where $m$ is fixed and $p$ is an unknown proportion.
(ii) $X_1 \sim Uniform[a, b]$, where $a$ and $b$ are unknown constants.
(iii) $X_1 \sim N(\mu, \sigma^2)$, where $\mu$ and $\sigma^2$ are unknown parameters.

2.15. Suppose that $X_1, X_2, \ldots$ are i.i.d. with a Cauchy($0, 1$) distribution; that is, the pdf of $X_i$ is given by
\[
f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty.
\]
Find a positive number $\delta$ such that $n^{-\delta}X_{(n)}$ converges in distribution to a nondegenerate distribution, where $X_{(n)} = \max_{1 \leq i \leq n} X_i$. What is the limiting distribution?

2.16. Suppose that $X_1, \ldots, X_n$ are i.i.d. Exponential(1) random variables. Define $X_{(n)}$ as in Exercise 2.15. Show that

$$X_{(n)} - \log(n) \xrightarrow{d} \xi$$

as $n \to \infty$, where the cdf of $\xi$ is given by

$$F(x) = \exp\{-\exp(-x)\}, \quad -\infty < x < \infty.$$

2.17. Let $X_1, X_2, \ldots$ be i.i.d. Uniform$(0,1]$ random variables and $\xi_n = (\prod_{i=1}^n X_i)^{-1/n}$. Show that

$$\sqrt{n}(\xi_n - e) \xrightarrow{d} \xi$$

as $n \to \infty$, where $\xi \sim \text{N}(0, e^2)$. (Hint: The result can be established as an application of the CLT; see Chapter 6.)

2.18. Complete the second half of the proof of Theorem 2.9; that is, $\limsup F_n(x) \leq F(x + \epsilon)$ for any $\epsilon > 0$.

2.19. Give examples of a random variable $\xi$ such that the following hold:

(i) The mgf of $\xi$ does not exist for any $t$ except $t = 0$.
(ii) The mgf of $\xi$ exists for $|t| < 1$ but does not exist for $|t| \geq 1$.
(iii) The mgf of $\xi$ exists for any $t \in \mathbb{R}$.

2.20. Show that the integrand in (2.10) is bounded in absolute value, and therefore the expectation exists for any $t \in \mathbb{R}$.

2.21. Suppose that $\xi_n \sim t_n$, $n = 1, 2, \ldots$. Show that the following hold:

(i) $\xi_n \xrightarrow{d} \xi \sim \text{N}(0,1)$.
(ii) $m_n(t) = E(e^{t\xi_n}) = \infty, \forall t \neq 0$.
(iii) $m(t) = E(e^{t\xi}) = e^{t^2/2}, t \in \mathbb{R}$.

2.22. Derive the results of Lemma 2.2.

2.23. Derive the results of Lemma 2.3.

2.24. (i) Suppose that $\xi \sim \text{Binomial}(n, p)$. Show that $m_{\xi}(t) = (pe^t + 1 - p)^n$.
(ii) Show that $(pn e^t + 1 - p)^n \to \exp\{\lambda(e^t - 1)\}$ as $n \to \infty$, $t \in \mathbb{R}$, provided that $npn \to \lambda$ as $n \to \infty$.

2.25. Suppose that $X_1, \ldots, X_n$ are i.i.d. with the pdf

$$f(x) = \frac{1 - \cos(x)}{\pi x^2}, \quad -\infty < x < \infty.$$

(i) Show that the mgf of $X_i$ does not exist.
(ii) Show that the cf of $X_i$ is given by $\max(1 - |t|, 0)$, $t \in \mathbb{R}$.
(iii) Show that the cf of $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ is given by

$$\left\{ \max \left( 1 - \frac{|t|}{n}, 0 \right) \right\}^n,$$
48 2 Modes of Convergence

which converges to $e^{-|t|}$ as $n \to \infty$.

(iv) Show that the cf of $\xi \sim \text{Cauchy}(0, 1)$ is $e^{-|t|}$, $t \in R$. Therefore, $\bar{X} \overset{d}{\to} \xi$ as $n \to \infty$.


2.27. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli($p$) observations. Show that

$$\left\{ \frac{n}{p(1-p)} \right\}^{1/2} (\hat{p} - p) \overset{d}{\to} N(0, 1) \text{ as } n \to \infty,$$

where $\hat{p}$ is the sample proportion which is equal to $(X_1 + \cdots + X_n)/n$. This result is also known as normal approximation to binomial distribution. (Of course, the result follows from the CLT, but here you are asked to show it directly—without using the CLT.)

2.28. Suppose that $\xi_n \overset{P}{\to} \xi$ as $n \to \infty$ and $g$ is a bounded continuous function. Show that $g(\xi_n) \overset{L^p}{\to} g(\xi)$ as $n \to \infty$ for every $p > 0$.

2.29. Let $X \sim \text{Uniform}(0, 1)$. Define $\xi_n = 2n^{-1}1(0 < X < 1/n)$, $n = 1, 2, \ldots$.

(i) Show that $\xi_n \overset{a.s.}{\to} 0$ as $n \to \infty$.

(ii) Show that $\xi_n$, $n = 1, 2, \ldots$, does not converge to zero in $L^p$ for any $p > 0$.

2.30. Prove Theorem 2.15 using Chebyshev’s inequality (see Section 5.2).

2.31. Use Skorokhod’s theorem (Theorem 2.18) to prove the first half of Theorem 2.11; that is, convergence in distribution implies convergence of the characteristic function.

2.32. Prove that the inequality (2.16) holds for any $p > 0$. Note that for $p \geq 1$, this follows from the convex function inequality, but the inequality holds for $0 < p < 1$ as well.

2.33. Complete the proof of Theorem 2.17 (i) $\Rightarrow$ (ii) using the argument of subsequences (see §1.5.1.6).

2.34. Use the dominated convergence theorem (Theorem 2.16) to show (2.20). Also show that $E(||\xi_n||^p - ||\xi||^p) \to 0$ implies $E(||\xi_n||^p) \to E(||\xi||^p)$ as $n \to 0$.

2.35. Refer to the (iii) $\Rightarrow$ (i) part of the proof of Theorem 2.17.

(i) Show that $\eta_n \overset{P}{\to} \eta$ as $n \to \infty$.

(ii) Show that it is not necessarily true that $|\xi_n|^p 1(|\xi_n| \leq a) \overset{P}{\to} |\xi|^p 1(|\xi| \leq a)$ as $n \to \infty$.

2.36. This exercise refers to Example 2.17.

(i) Show that $X_1, \ldots, X_n$ are i.i.d. with a distribution whose cf is given by (2.25).

(ii) If we define $X_i(\theta) = \bar{Y}_i - \bar{x}_i \beta$ for an arbitrary $\theta = \beta$ (not necessarily the true parameter vector), then conditions (i) and (ii) are satisfied.

2.37. Consider the function $f(x, y) = x^2 + y^2$. Show that

$$\left. \frac{\partial}{\partial x} f(x, y) \right|_{y=x} \neq \frac{\partial}{\partial x} f(x, x).$$
2.38. Show that (2.33) is equivalent to that (2.35) holds for every $\lambda \in R^M$.

2.39. Regarding the distribution of $|\xi|^2 = \xi'\xi$ in (2.37), show the following [see the notation below (2.37)]:

(i) $PB = BP = 0$.

(ii) $P$ is a projection matrix with rank 1.

(iii) The distribution of $\xi'\xi$ is the same as (2.29), where $Z_1, \ldots, Z_{M-1}$ are independent standard normal random variables.

2.40. Use the Borel–Cantelli lemma (Lemma 2.5) to prove the following:

(i) If for every $\epsilon > 0$ we have $\sum_{n=1}^{\infty} P(|\xi_n - \xi| \geq \epsilon) < \infty$, then $\xi_n \xrightarrow{a.s.} \xi$ as $n \to \infty$.

(ii) If $\sum_{n=1}^{\infty} E(|\xi_n - \xi|^p) < \infty$ for some $p > 0$, then $\xi_n \xrightarrow{a.s.} \xi$ as $n \to \infty$. 
Large Sample Techniques for Statistics
Jiang, J.
2010, XVIII, 610 p., Hardcover