2

Stochastic Convergence

This chapter provides a review of basic modes of convergence of sequences of stochastic vectors, in particular convergence in distribution and in probability.

2.1 Basic Theory

A *random vector* in $\mathbb{R}^k$ is a vector $X = (X_1, \dots, X_k)$ of real random variables.† The *distribution function* of X is the map $x \mapsto \mathrm{P}(X \leq x)$.

A sequence of random vectors X_n is said to *converge in distribution* to a random vector X if

$$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x),$$

for every x at which the limit distribution function $x \mapsto \mathrm{P}(X \leq x)$ is continuous. Alternative names are *weak convergence* and *convergence in law*. As the last name suggests, the convergence only depends on the induced laws of the vectors and not on the probability spaces on which they are defined. Weak convergence is denoted by $X_n \rightsquigarrow X$; if X has distribution L, or a distribution with a standard code, such as $N(0, 1)$, then also by $X_n \rightsquigarrow L$ or $X_n \rightsquigarrow N(0, 1)$.

Let $d(x, y)$ be a distance function on $\mathbb{R}^k$ that generates the usual topology. For instance, the Euclidean distance

$$d(x, y) = \|x - y\| = \left(\sum_{i=1}^{k} (x_i - y_i)^2\right)^{1/2}.$$

A sequence of random variables X_n is said to *converge in probability* to X if for all $\varepsilon > 0$

$$\mathrm{P}\big(d(X_n, X) > \varepsilon\big) \to 0.$$

This is denoted by $X_n \xrightarrow{\mathrm{P}} X$. In this notation convergence in probability is the same as $d(X_n, X) \xrightarrow{\mathrm{P}} 0$.

† More formally it is a Borel measurable map from some probability space in $\mathbb{R}^k$. Throughout it is implicitly understood that variables X, $g(X)$, and so forth of which we compute expectations or probabilities are measurable maps on some probability space.

As we shall see, convergence in probability is stronger than convergence in distribution. An even stronger mode of convergence is almost-sure convergence. The sequence X_n is said to *converge almost surely* to X if $d(X_n, X) \to 0$ with probability one:

$$\mathrm{P}\big(\lim d(X_n, X) = 0\big) = 1.$$

This is denoted by $X_n \xrightarrow{\text{as}} X$. Note that convergence in probability and convergence almost surely only make sense if each of X_n and X are defined on the same probability space. For convergence in distribution this is not necessary.

2.1 ***Example (Classical limit theorems).*** Let $\overline{Y}_n$ be the average of the first n of a sequence of independent, identically distributed random vectors $Y_1, Y_2, \dots$. If $\mathrm{E}\|Y_1\| < \infty$, then $\overline{Y}_n \xrightarrow{\text{as}} \mathrm{E}Y_1$ by the *strong law of large numbers*. Under the stronger assumption that $\mathrm{E}\|Y_1\|^2 < \infty$, the *central limit theorem* asserts that $\sqrt{n}(\overline{Y}_n - \mathrm{E}Y_1) \rightsquigarrow N(0, \mathrm{Cov}\, Y_1)$. The central limit theorem plays an important role in this manuscript. It is proved later in this chapter, first for the case of real variables, and next it is extended to random vectors. The strong law of large numbers appears to be of less interest in statistics. Usually the *weak law of large numbers*, according to which $\overline{Y}_n \xrightarrow{\mathrm{P}} \mathrm{E}Y_1$, suffices. This is proved later in this chapter. □

The portmanteau lemma gives a number of equivalent descriptions of weak convergence. Most of the characterizations are only useful in proofs. The last one also has intuitive value.

2.2 ***Lemma (Portmanteau).*** *For any random vectors X_n and X the following statements are equivalent.*

(i) $\mathrm{P}(X_n \le x) \to \mathrm{P}(X \le x)$ *for all continuity points of* $x \mapsto \mathrm{P}(X \le x)$;
(ii) $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for all bounded, continuous functions* f;
(iii) $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for all bounded, Lipschitz*[†] *functions* f;
(iv) $\liminf \mathrm{E}f(X_n) \ge \mathrm{E}f(X)$ *for all nonnegative, continuous functions* f;
(v) $\liminf \mathrm{P}(X_n \in G) \ge \mathrm{P}(X \in G)$ *for every open set* G;
(vi) $\limsup \mathrm{P}(X_n \in F) \le \mathrm{P}(X \in F)$ *for every closed set* F;
(vii) $\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$ *for all Borel sets* B *with* $\mathrm{P}(X \in \delta B) = 0$, *where* $\delta B = \overline{B} - \mathring{B}$ *is the boundary of* B.

Proof. (i) $\Rightarrow$ (ii). Assume first that the distribution function of X is continuous. Then condition (i) implies that $\mathrm{P}(X_n \in I) \to \mathrm{P}(X \in I)$ for every rectangle I. Choose a sufficiently large, compact rectangle I with $\mathrm{P}(X \notin I) < \varepsilon$. A continuous function f is uniformly continuous on the compact set I. Thus there exists a partition $I = \cup_j I_j$ into finitely many rectangles I_j such that f varies at most ε on every I_j. Take a point x_j from each I_j and define $f_\varepsilon = \sum_j f(x_j) 1_{I_j}$. Then $|f - f_\varepsilon| < \varepsilon$ on I, whence if f takes its values in $[-1, 1]$,

$$\begin{aligned} \big|\mathrm{E}f(X_n) - \mathrm{E}f_\varepsilon(X_n)\big| &\le \varepsilon + \mathrm{P}(X_n \notin I), \\ \big|\mathrm{E}f(X) - \mathrm{E}f_\varepsilon(X)\big| &\le \varepsilon + \mathrm{P}(X \notin I) < 2\varepsilon. \end{aligned}$$

[†] A function is called *Lipschitz* if there exists a number L such that $\big|f(x) - f(y)\big| \le L d(x, y)$, for every x and y. The least such number L is denoted $\|f\|_{\mathrm{lip}}$.

For sufficiently large n, the right side of the first equation is smaller than 2ε as well. We combine this with

$$\left|\mathrm{E}f_\varepsilon(X_n) - \mathrm{E}f_\varepsilon(X)\right| \leq \sum_j \left|\mathrm{P}(X_n \in I_j) - \mathrm{P}(X \in I_j)\right|\left|f(x_j)\right| \to 0.$$

Together with the triangle inequality the three displays show that $\left|\mathrm{E}f(X_n) - \mathrm{E}f(X)\right|$ is bounded by 5ε eventually. This being true for every $\varepsilon > 0$ implies (ii).

Call a set B a *continuity set* if its boundary δB satisfies $\mathrm{P}(X \in \delta B) = 0$. The preceding argument is valid for a general X provided all rectangles I are chosen equal to continuity sets. This is possible, because the collection of discontinuity sets is sparse. Given any collection of pairwise disjoint measurable sets, at most countably many sets can have positive probability. Otherwise the probability of their union would be infinite. Therefore, given any collection of sets $\{B_\alpha : \alpha \in A\}$ with pairwise disjoint boundaries, all except at most countably many sets are continuity sets. In particular, for each j at most countably many sets of the form $\{x : x_j \leq \alpha\}$ are not continuity sets. Conclude that there exist dense subsets $Q_1, \dots, Q_k$ of $\mathbb{R}$ such that each rectangle with corners in the set $Q_1 \times \cdots \times Q_k$ is a continuity set. We can choose all rectangles I inside this set.

(iii) $\Rightarrow$ (v). For every open set G there exists a sequence of Lipschitz functions with $0 \leq f_m \uparrow 1_G$. For instance $f_m(x) = \bigl(md(x, G^c)\bigr) \wedge 1$. For every fixed m,

$$\liminf_{n\to\infty} \mathrm{P}(X_n \in G) \geq \liminf_{n\to\infty} \mathrm{E}f_m(X_n) = \mathrm{E}f_m(X).$$

As $m \to \infty$ the right side increases to $\mathrm{P}(X \in G)$ by the monotone convergence theorem.

(v) $\Leftrightarrow$ (vi). Because a set is open if and only if its complement is closed, this follows by taking complements.

(v) + (vi) $\Rightarrow$ (vii). Let $\mathring{B}$ and $\overline{B}$ denote the interior and the closure of a set, respectively. By (iv)

$$\mathrm{P}(X \in \mathring{B}) \leq \liminf \mathrm{P}(X_n \in \mathring{B}) \leq \limsup \mathrm{P}(X_n \in \overline{B}) \leq \mathrm{P}(X \in \overline{B}),$$

by (v). If $\mathrm{P}(X \in \delta B) = 0$, then left and right side are equal, whence all inequalities are equalities. The probability $\mathrm{P}(X \in B)$ and the limit $\lim \mathrm{P}(X_n \in B)$ are between the expressions on left and right and hence equal to the common value.

(vii) $\Rightarrow$ (i). Every cell $(-\infty, x]$ such that x is a continuity point of $x \mapsto \mathrm{P}(X \leq x)$ is a continuity set.

The equivalence (ii) $\Leftrightarrow$ (iv) is left as an exercise. ■

The continuous-mapping theorem is a simple result, but it is extremely useful. If the sequence of random vectors X_n converges to X and g is continuous, then $g(X_n)$ converges to $g(X)$. This is true for each of the three modes of stochastic convergence.

2.3 *Theorem (Continuous mapping).* *Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set C such that* $\mathrm{P}(X \in C) = 1$.

(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$;
(ii) If $X_n \xrightarrow{\mathrm{P}} X$, then $g(X_n) \xrightarrow{\mathrm{P}} g(X)$;
(iii) If $X_n \xrightarrow{\mathrm{as}} X$, then $g(X_n) \xrightarrow{\mathrm{as}} g(X)$.

Proof. (i). The event $\{g(X_n) \in F\}$ is identical to the event $\{X_n \in g^{-1}(F)\}$. For every closed set F,

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset g^{-1}(F) \cup C^c.$$

To see the second inclusion, take x in the closure of $g^{-1}(F)$. Thus, there exists a sequence x_m with $x_m \to x$ and $g(x_m) \in F$ for every F. If $x \in C$, then $g(x_m) \to g(x)$, which is in F because F is closed; otherwise $x \in C^c$. By the portmanteau lemma,

$$\limsup \mathrm{P}\big(g(X_n) \in F\big) \leq \limsup \mathrm{P}\big(X_n \in \overline{g^{-1}(F)}\big) \leq \mathrm{P}\big(X \in \overline{g^{-1}(F)}\big).$$

Because $\mathrm{P}(X \in C^c) = 0$, the probability on the right is $\mathrm{P}\big(X \in g^{-1}(F)\big) = \mathrm{P}\big(g(X) \in F\big)$. Apply the portmanteau lemma again, in the opposite direction, to conclude that $g(X_n) \rightsquigarrow g(X)$.

(ii). Fix arbitrary $\varepsilon > 0$. For each $\delta > 0$ let B_δ be the set of x for which there exists y with $d(x, y) < \delta$, but $d\big(g(x), g(y)\big) > \varepsilon$. If $X \notin B_\delta$ and $d\big(g(X_n), g(X)\big) > \varepsilon$, then $d(X_n, X) \geq \delta$. Consequently,

$$\mathrm{P}\Big(d\big(g(X_n), g(X)\big) > \varepsilon\Big) \leq \mathrm{P}(X \in B_\delta) + \mathrm{P}\big(d(X_n, X) \geq \delta\big).$$

The second term on the right converges to zero as $n \to \infty$ for every fixed $\delta > 0$. Because $B_\delta \cap C \downarrow \emptyset$ by continuity of g, the first term converges to zero as $\delta \downarrow 0$.

Assertion (iii) is trivial. ∎

Any random vector X is *tight*: For every $\varepsilon > 0$ there exists a constant M such that $\mathrm{P}\big(\|X\| > M\big) < \varepsilon$. A set of random vectors $\{X_\alpha : \alpha \in A\}$ is called *uniformly tight* if M can be chosen the same for every X_α: For every $\varepsilon > 0$ there exists a constant M such that

$$\sup_\alpha \mathrm{P}\big(\|X_\alpha\| > M\big) < \varepsilon.$$

Thus, there exists a compact set to which all X_α give probability "almost" one. Another name for uniformly tight is *bounded in probability*. It is not hard to see that every weakly converging sequence X_n is uniformly tight. More surprisingly, the converse of this statement is almost true: According to Prohorov's theorem, every uniformly tight sequence contains a weakly converging subsequence. Prohorov's theorem generalizes the Heine-Borel theorem from deterministic sequences X_n to random vectors.

2.4 ***Theorem (Prohorov's theorem).*** *Let X_n be random vectors in $\mathbb{R}^k$.*

(i) If $X_n \rightsquigarrow X$ for some X, then $\{X_n : n \in \mathbb{N}\}$ is uniformly tight;

(ii) If X_n is uniformly tight, then there exists a subsequence with $X_{n_j} \rightsquigarrow X$ as $j \to \infty$, for some X.

Proof. (i). Fix a number M such that $\mathrm{P}\big(\|X\| \geq M\big) < \varepsilon$. By the portmanteau lemma $\mathrm{P}\big(\|X_n\| \geq M\big)$ exceeds $\mathrm{P}\big(\|X\| \geq M\big)$ arbitrarily little for sufficiently large n. Thus there exists N such that $\mathrm{P}\big(\|X_n\| \geq M\big) < 2\varepsilon$, for all $n \geq N$. Because each of the finitely many variables X_n with $n < N$ is tight, the value of M can be increased, if necessary, to ensure that $\mathrm{P}\big(\|X_n\| \geq M\big) < 2\varepsilon$ for every n.

(ii). By Helly's lemma (described subsequently), there exists a subsequence F_{n_j} of the sequence of cumulative distribution functions $F_n(x) = \mathrm{P}(X_n \leq x)$ that converges weakly to a possibly "defective" distribution function F. It suffices to show that F is a proper distribution function: $F(x) \to 0, 1$ if $x_i \to -\infty$ for some i, or $x \to \infty$. By the uniform tightness, there exists M such that $F_n(M) > 1-\varepsilon$ for all n. By making M larger, if necessary, it can be ensured that M is a continuity point of F. Then $F(M) = \lim F_{n_j}(M) \geq 1-\varepsilon$. Conclude that $F(x) \to 1$ as $x \to \infty$. That the limits at $-\infty$ are zero can be seen in a similar manner. ∎

The crux of the proof of Prohorov's theorem is Helly's lemma. This asserts that any given sequence of distribution functions contains a subsequence that converges weakly to a possibly defective distribution function. A *defective distribution function* is a function that has all the properties of a cumulative distribution function with the exception that it has limits less than 1 at ∞ and/or greater than 0 at $-\infty$.

2.5 Lemma (Helly's lemma). *Each given sequence F_n of cumulative distribution functions on $\mathbb{R}^k$ possesses a subsequence F_{n_j} with the property that $F_{n_j}(x) \to F(x)$ at each continuity point x of a possibly defective distribution function F.*

Proof. Let $\mathbb{Q}^k = \{q_1, q_2, \ldots\}$ be the vectors with rational coordinates, ordered in an arbitrary manner. Because the sequence $F_n(q_1)$ is contained in the interval $[0, 1]$, it has a converging subsequence. Call the indexing subsequence $\{n_j^1\}_{j=1}^\infty$ and the limit $G(q_1)$. Next, extract a further subsequence $\{n_j^2\} \subset \{n_j^1\}$ along which $F_n(q_2)$ converges to a limit $G(q_2)$, a further subsequence $\{n_j^3\} \subset \{n_j^2\}$ along which $F_n(q_3)$ converges to a limit $G(q_3), \ldots,$ and so forth. The "tail" of the diagonal sequence $n_j := n_j^j$ belongs to every sequence n_j^i. Hence $F_{n_j}(q_i) \to G(q_i)$ for every $i = 1, 2, \ldots$. Because each F_n is nondecreasing, $G(q) \leq G(q')$ if $q \leq q'$. Define

$$F(x) = \inf_{q > x} G(q).$$

Then F is nondecreasing. It is also right-continuous at every point x, because for every $\varepsilon > 0$ there exists $q > x$ with $G(q) - F(x) < \varepsilon$, which implies $F(y) - F(x) < \varepsilon$ for every $x \leq y \leq q$. Continuity of F at x implies, for every $\varepsilon > 0$, the existence of $q < x < q'$ such that $G(q') - G(q) < \varepsilon$. By monotonicity, we have $G(q) \leq F(x) \leq G(q')$, and

$$G(q) = \lim F_{n_j}(q) \leq \liminf F_{n_j}(x) \leq \lim F_{n_j}(q') = G(q').$$

Conclude that $\left|\liminf F_{n_j}(x) - F(x)\right| < \varepsilon$. Because this is true for every $\varepsilon > 0$ and the same result can be obtained for the lim sup, it follows that $F_{n_j}(x) \to F(x)$ at every continuity point of F.

In the higher-dimensional case, it must still be shown that the expressions defining masses of cells are nonnegative. For instance, for $k = 2$, F is a (defective) distribution function only if $F(b) + F(a) - F(a_1, b_2) - F(a_2, b_1) \geq 0$ for every $a \leq b$. In the case that the four corners a, b, (a_1, b_2), and (a_2, b_1) of the cell are continuity points; this is immediate from the convergence of F_{n_j} to F and the fact that each F_n is a distribution function. Next, for general cells the property follows by right continuity. ∎

2.6 ***Example (Markov's inequality).*** A sequence X_n of random variables with $\mathrm{E}|X_n|^p = O(1)$ for some $p > 0$ is uniformly tight. This follows because by *Markov's inequality*

$$\mathrm{P}\big(|X_n| > M\big) \leq \frac{\mathrm{E}|X_n|^p}{M^p}.$$

The right side can be made arbitrarily small, uniformly in n, by choosing sufficiently large M.

Because $\mathrm{E}X_n^2 = \mathrm{var}\, X_n + (\mathrm{E}X_n)^2$, an alternative sufficient condition for uniform tightness is $\mathrm{E}X_n = O(1)$ and $\mathrm{var}\, X_n = O(1)$. This cannot be reversed. □

Consider some of the relationships among the three modes of convergence. Convergence in distribution is weaker than convergence in probability, which is in turn weaker than almost-sure convergence, except if the limit is constant.

2.7 ***Theorem.*** *Let X_n, X and Y_n be random vectors. Then*

(i) $X_n \xrightarrow{\text{as}} X$ implies $X_n \xrightarrow{\mathrm{P}} X$;
(ii) $X_n \xrightarrow{\mathrm{P}} X$ implies $X_n \rightsquigarrow X$;
(iii) $X_n \xrightarrow{\mathrm{P}} c$ for a constant c if and only if $X_n \rightsquigarrow c$;
(iv) if $X_n \rightsquigarrow X$ and $d(X_n, Y_n) \xrightarrow{\mathrm{P}} 0$, then $Y_n \rightsquigarrow X$;
(v) if $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{\mathrm{P}} c$ for a constant c, then $(X_n, Y_n) \rightsquigarrow (X, c)$;
(vi) if $X_n \xrightarrow{\mathrm{P}} X$ and $Y_n \xrightarrow{\mathrm{P}} Y$, then $(X_n, Y_n) \xrightarrow{\mathrm{P}} (X, Y)$.

Proof. (i). The sequence of sets $A_n = \cup_{m \geq n}\big\{d(X_m, X) > \varepsilon\big\}$ is decreasing for every $\varepsilon > 0$ and decreases to the empty set if $X_n(\omega) \to X(\omega)$ for every ω. If $X_n \xrightarrow{\text{as}} X$, then $\mathrm{P}\big(d(X_n, X) > \varepsilon\big) \leq \mathrm{P}(A_n) \to 0$.

(iv). For every f with range $[0, 1]$ and Lipschitz norm at most 1 and every $\varepsilon > 0$,

$$\big|\mathrm{E}f(X_n) - \mathrm{E}f(Y_n)\big| \leq \varepsilon \mathrm{E}1\big\{d(X_n, Y_n) \leq \varepsilon\big\} + 2\mathrm{E}1\big\{d(X_n, Y_n) > \varepsilon\big\}.$$

The second term on the right converges to zero as $n \to \infty$. The first term can be made arbitrarily small by choice of ε. Conclude that the sequences $\mathrm{E}f(X_n)$ and $\mathrm{E}f(Y_n)$ have the same limit. The result follows from the portmanteau lemma.

(ii). Because $d(X_n, X) \xrightarrow{\mathrm{P}} 0$ and trivially $X \rightsquigarrow X$, it follows that $X_n \rightsquigarrow X$ by (iv).

(iii). The "only if" part is a special case of (ii). For the converse let $\mathrm{ball}(c, \varepsilon)$ be the open ball of radius ε around c. Then $\mathrm{P}\big(d(X_n, c) \geq \varepsilon\big) = \mathrm{P}\big(X_n \in \mathrm{ball}(c, \varepsilon)^c\big)$. If $X_n \rightsquigarrow c$, then the lim sup of the last probability is bounded by $\mathrm{P}\big(c \in \mathrm{ball}(c, \varepsilon)^c\big) = 0$, by the portmanteau lemma.

(v). First note that $d\big((X_n, Y_n), (X_n, c)\big) = d(Y_n, c) \xrightarrow{\mathrm{P}} 0$. Thus, according to (iv), it suffices to show that $(X_n, c) \rightsquigarrow (X, c)$. For every continuous, bounded function $(x, y) \mapsto f(x, y)$, the function $x \mapsto f(x, c)$ is continuous and bounded. Thus $\mathrm{E}f(X_n, c) \to \mathrm{E}f(X, c)$ if $X_n \rightsquigarrow X$.

(vi). This follows from $d\big((x_1, y_1), (x_2, y_2)\big) \leq d(x_1, x_2) + d(y_1, y_2)$. ■

According to the last assertion of the lemma, convergence in probability of a sequence of vectors $X_n = (X_{n,1}, \ldots, X_{n,k})$ is equivalent to convergence of every one of the sequences of components $X_{n,i}$ separately. The analogous statement for convergence in distribution is false: Convergence in distribution of the sequence X_n is stronger than convergence of every one of the sequences of components $X_{n,i}$. The point is that the distribution of the components $X_{n,i}$ separately does not determine their joint distribution: They might be independent or dependent in many ways. We speak of *joint convergence* in distribution versus *marginal convergence* .

Assertion (v) of the lemma has some useful consequences. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $(X_n, Y_n) \rightsquigarrow (X, c)$. Consequently, by the continuous mapping theorem, $g(X_n, Y_n) \rightsquigarrow g(X, c)$ for every map g that is continuous at every point in the set $\mathbb{R}^k \times \{c\}$ in which the vector (X, c) takes its values. Thus, for every g such that

$$\lim_{x \to x_0, y \to c} g(x, y) = g(x_0, c), \qquad \text{for every } x_0.$$

Some particular applications of this principle are known as Slutsky's lemma.

2.8 ***Lemma (Slutsky).*** *Let X_n, X and Y_n be random vectors or variables. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ for a constant c, then*

(i) $X_n + Y_n \rightsquigarrow X + c$;
(ii) $Y_n X_n \rightsquigarrow cX$;
(iii) $Y_n^{-1} X_n \rightsquigarrow c^{-1} X$ *provided* $c \neq 0$.

In (i) the "constant" c must be a vector of the same dimension as X, and in (ii) it is probably initially understood to be a scalar. However, (ii) is also true if every Y_n and c are matrices (which can be identified with vectors, for instance by aligning rows, to give a meaning to the convergence $Y_n \rightsquigarrow c$), simply because matrix multiplication $(x, y) \mapsto yx$ is a continuous operation. Even (iii) is valid for matrices Y_n and c and vectors X_n provided $c \neq 0$ is understood as c being invertible, because taking an inverse is also continuous.

2.9 ***Example (t-statistic).*** Let $Y_1, Y_2, \ldots$ be independent, identically distributed random variables with $\mathrm{E}Y_1 = 0$ and $\mathrm{E}Y_1^2 < \infty$. Then the t-statistic $\sqrt{n}\,\overline{Y}_n / S_n$, where $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ is the sample variance, is asymptotically standard normal.

To see this, first note that by two applications of the weak law of large numbers and the continuous-mapping theorem for convergence in probability

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \overline{Y}_n^2\right) \xrightarrow{\mathrm{P}} 1\big(\mathrm{E}Y_1^2 - (\mathrm{E}Y_1)^2\big) = \mathrm{var}\, Y_1.$$

Again by the continuous-mapping theorem, S_n converges in probability to $\mathrm{sd}\, Y_1$. By the central limit theorem $\sqrt{n}\,\overline{Y}_n$ converges in law to the $N(0, \mathrm{var}\, Y_1)$ distribution. Finally, Slutsky's lemma gives that the sequence of t-statistics converges in distribution to $N(0, \mathrm{var}\, Y_1)/\, \mathrm{sd}\, Y_1 = N(0, 1)$. □

2.10 ***Example (Confidence intervals).*** Let T_n and S_n be sequences of estimators satisfying

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, \sigma^2), \qquad S_n^2 \xrightarrow{\mathrm{P}} \sigma^2,$$

for certain parameters θ and σ^2 depending on the underlying distribution, for every distribution in the model. Then $\theta = T_n \pm S_n/\sqrt{n}\, z_\alpha$ is a confidence interval for θ of asymptotic level $1 - 2\alpha$. More precisely, we have that the probability that θ is contained in $[T_n - S_n/\sqrt{n}\, z_\alpha, T_n + S_n/\sqrt{n}\, z_\alpha]$ converges to $1 - 2\alpha$.

This is a consequence of the fact that the sequence $\sqrt{n}(T_n - \theta)/S_n$ is asymptotically standard normally distributed. □

If the limit variable X has a continuous distribution function, then weak convergence $X_n \rightsquigarrow X$ implies $\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x)$ for every x. The convergence is then even uniform in x.

2.11 ***Lemma.*** *Suppose that $X_n \rightsquigarrow X$ for a random vector X with a continuous distribution function. Then* $\sup_x \bigl|\mathrm{P}(X_n \leq x) - \mathrm{P}(X \leq x)\bigr| \to 0$.

Proof. Let F_n and F be the distribution functions of X_n and X. First consider the one-dimensional case. Fix $k \in \mathbb{N}$. By the continuity of F there exist points $-\infty = x_0 < x_1 < \cdots < x_k = \infty$ with $F(x_i) = i/k$. By monotonicity, we have, for $x_{i-1} \leq x \leq x_i$,

$$\begin{aligned} F_n(x) - F(x) &\leq F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + 1/k \\ &\geq F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - 1/k. \end{aligned}$$

Thus $\bigl|F_n(x) - F(x)\bigr|$ is bounded above by $\sup_i \bigl|F_n(x_i) - F(x_i)\bigr| + 1/k$, for every x. The latter, finite supremum converges to zero as $n \to \infty$, for each fixed k. Because k is arbitrary, the result follows.

In the higher-dimensional case, we follow a similar argument but use hyperrectangles, rather than intervals. We can construct the rectangles by intersecting the k partitions obtained by subdividing each coordinate separately as before. ■

2.2 Stochastic o and O Symbols

It is convenient to have short expressions for terms that converge in probability to zero or are uniformly tight. The notation $o_P(1)$ ("small oh-P-one") is short for a sequence of random vectors that converges to zero in probability. The expression $O_P(1)$ ("big oh-P-one") denotes a sequence that is bounded in probability. More generally, for a given sequence of random variables R_n,

$$\begin{aligned} X_n &= o_P(R_n) &\quad \text{means} \quad & X_n = Y_n R_n \quad \text{and} \quad Y_n \xrightarrow{\mathrm{P}} 0; \\ X_n &= O_P(R_n) &\quad \text{means} \quad & X_n = Y_n R_n \quad \text{and} \quad Y_n = O_P(1). \end{aligned}$$

This expresses that the sequence X_n converges in probability to zero or is bounded in probability at the "rate" R_n. For deterministic sequences X_n and R_n, the stochastic "oh" symbols reduce to the usual o and O from calculus.

There are many rules of calculus with o and O symbols, which we apply without comment. For instance,

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ O_P(1) o_P(1) &= o_P(1) \end{aligned}$$

$$
\begin{aligned}
\big(1 + o_P(1)\big)^{-1} &= O_P(1) \\
o_P(R_n) &= R_n o_P(1) \\
O_P(R_n) &= R_n O_P(1) \\
o_P\big(O_P(1)\big) &= o_P(1).
\end{aligned}
$$

To see the validity of these rules it suffices to restate them in terms of explicitly named vectors, where each $o_P(1)$ and $O_P(1)$ should be replaced by a different sequence of vectors that converges to zero or is bounded in probability. In this way the first rule says: If $X_n \xrightarrow{\mathrm{P}} 0$ and $Y_n \xrightarrow{\mathrm{P}} 0$, then $Z_n = X_n + Y_n \xrightarrow{\mathrm{P}} 0$. This is an example of the continuous-mapping theorem. The third rule is short for the following: If X_n is bounded in probability and $Y_n \xrightarrow{\mathrm{P}} 0$, then $X_n Y_n \xrightarrow{\mathrm{P}} 0$. If X_n would also converge in distribution, then this would be statement (ii) of Slutsky's lemma (with $c = 0$). But by Prohorov's theorem, X_n converges in distribution "along subsequences" if it is bounded in probability, so that the third rule can still be deduced from Slutsky's lemma by "arguing along subsequences."

Note that both rules are in fact implications and should be read from left to right, even though they are stated with the help of the equality sign. Similarly, although it is true that $o_P(1) + o_P(1) = 2o_P(1)$, writing down this rule does not reflect understanding of the o_P symbol.

Two more complicated rules are given by the following lemma.

2.12 ***Lemma.*** *Let R be a function defined on domain in $\mathbb{R}^k$ such that $R(0) = 0$. Let X_n be a sequence of random vectors with values in the domain of R that converges in probability to zero. Then, for every $p > 0$,*

(i) if $R(h) = o\big(\|h\|^p\big)$ as $h \to 0$, then $R(X_n) = o_P\big(\|X_n\|^p\big)$;

(ii) if $R(h) = O\big(\|h\|^p\big)$ as $h \to 0$, then $R(X_n) = O_P\big(\|X_n\|^p\big)$.

Proof. Define $g(h)$ as $g(h) = R(h)/\|h\|^p$ for $h \neq 0$ and $g(0) = 0$. Then $R(X_n) = g(X_n)\|X_n\|^p$.

(i) Because the function g is continuous at zero by assumption, $g(X_n) \xrightarrow{\mathrm{P}} g(0) = 0$ by the continuous-mapping theorem.

(ii) By assumption there exist M and $\delta > 0$ such that $|g(h)| \leq M$ whenever $\|h\| \leq \delta$. Thus $\mathrm{P}\big(|g(X_n)| > M\big) \leq \mathrm{P}\big(\|X_n\| > \delta\big) \to 0$, and the sequence $g(X_n)$ is tight. ■

*2.3 Characteristic Functions

It is sometimes possible to show convergence in distribution of a sequence of random vectors directly from the definition. In other cases "transforms" of probability measures may help. The basic idea is that it suffices to show characterization (ii) of the portmanteau lemma for a small subset of functions f only.

The most important transform is the *characteristic function*

$$t \mapsto \mathrm{E}e^{it^T X}, \qquad t \in \mathbb{R}^k.$$

Each of the functions $x \mapsto e^{it^T x}$ is continuous and bounded. Thus, by the portmanteau lemma, $\mathrm{E}e^{it^T X_n} \to \mathrm{E}e^{it^T X}$ for every t if $X_n \rightsquigarrow X$. By Lévy's continuity theorem the converse is also true: Pointwise convergence of characteristic functions is equivalent to weak convergence.

2.13 ***Theorem (Lévy's continuity theorem).*** *Let X_n and X be random vectors in $\mathbb{R}^k$. Then $X_n \rightsquigarrow X$ if and only if $\mathrm{E}e^{it^TX_n} \to \mathrm{E}e^{it^TX}$ for every $t \in \mathbb{R}^k$. Moreover, if $\mathrm{E}e^{it^TX_n}$ converges pointwise to a function $\phi(t)$ that is continuous at zero, then ϕ is the characteristic function of a random vector X and $X_n \rightsquigarrow X$.*

Proof. If $X_n \rightsquigarrow X$, then $\mathrm{E}h(X_n) \to \mathrm{E}h(X)$ for every bounded continuous function h, in particular for the functions $h(x) = e^{it^Tx}$. This gives one direction of the first statement.

For the proof of the last statement, suppose first that we already know that the sequence X_n is uniformly tight. Then, according to Prohorov's theorem, every subsequence has a further subsequence that converges in distribution to some vector Y. By the preceding paragraph, the characteristic function of Y is the limit of the characteristic functions of the converging subsequence. By assumption, this limit is the function $\phi(t)$. Conclude that every weak limit point Y of a converging subsequence possesses characteristic function ϕ. Because a characteristic function uniquely determines a distribution (see Lemma 2.15), it follows that the sequence X_n has only one weak limit point. It can be checked that a uniformly tight sequence with a unique limit point converges to this limit point, and the proof is complete.

The uniform tightness of the sequence X_n can be derived from the continuity of ϕ at zero. Because marginal tightness implies joint tightness, it may be assumed without loss of generality that X_n is one-dimensional. For every x and $\delta > 0$,

$$1\bigl\{|\delta x| > 2\bigr\} \le 2\left(1 - \frac{\sin \delta x}{\delta x}\right) = \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \cos tx)\,dt.$$

Replace x by X_n, take expectations, and use Fubini's theorem to obtain that

$$\mathrm{P}\left(|X_n| > \frac{2}{\delta}\right) \le \frac{1}{\delta}\int_{-\delta}^{\delta}\mathrm{Re}(1 - \mathrm{E}e^{itX_n})\,dt.$$

By assumption, the integrand in the right side converges pointwise to $\mathrm{Re}\bigl(1 - \phi(t)\bigr)$. By the dominated-convergence theorem, the whole expression converges to

$$\frac{1}{\delta}\int_{-\delta}^{\delta}\mathrm{Re}\bigl(1 - \phi(t)\bigr)\,dt.$$

Because ϕ is continuous at zero, there exists for every $\varepsilon > 0$ a $\delta > 0$ such that $\bigl|1 - \phi(t)\bigr| < \varepsilon$ for $|t| < \delta$. For this δ the integral is bounded by 2ε. Conclude that $\mathrm{P}\bigl(|X_n| > 2/\delta\bigr) \le 2\varepsilon$ for sufficiently large n, whence the sequence X_n is uniformly tight. ∎

2.14 ***Example (Normal distribution).*** The characteristic function of the $N_k(\mu, \Sigma)$ distribution is the function

$$t \mapsto e^{it^T\mu - \frac{1}{2}t^T\Sigma t}.$$

Indeed, if X is $N_k(0, I)$ distributed and $\Sigma^{1/2}$ is a symmetric square root of Σ (hence $\Sigma = (\Sigma^{1/2})^2$), then $\Sigma^{1/2}X + \mu$ possesses the given normal distribution and

$$\mathrm{E}e^{z^T(\Sigma^{1/2}X+\mu)} = e^{z^T\mu}\int e^{(\Sigma^{1/2}z)^Tx - \frac{1}{2}x^Tx}\,dx\,\frac{1}{(2\pi)^{k/2}} = e^{z^T\mu + \frac{1}{2}z^T\Sigma z}.$$

For real-valued z, the last equality follows easily by completing the square in the exponent. Evaluating the integral for complex z, such as $z = it$, requires some skill in complex function theory. One method, which avoids further calculations, is to show that both the left- and righthand sides of the preceding display are analytic functions of z. For the right side this is obvious; for the left side we can justify differentiation under the expectation sign by the dominated-convergence theorem. Because the two sides agree on the real axis, they must agree on the complex plane by uniqueness of analytic continuation. □

2.15 ***Lemma.*** *Random vectors X and Y in $\mathbb{R}^k$ are equal in distribution if and only if $\mathrm{E}e^{it^TX} = \mathrm{E}e^{it^TY}$ for every $t \in \mathbb{R}^k$.*

Proof. By Fubini's theorem and calculations as in the preceding example, for every $\sigma > 0$ and $y \in \mathbb{R}^k$,

$$\int e^{-it^Ty}\, e^{-\frac{1}{2}t^Tt\sigma^2}\, \mathrm{E}e^{it^TX}\, dt = \mathrm{E}\int e^{it^T(X-y)}\, e^{-\frac{1}{2}t^Tt\sigma^2}\, dt$$
$$= \frac{(2\pi)^{k/2}}{\sigma^k}\mathrm{E}e^{-\frac{1}{2}(X-y)^T(X-y)/\sigma^2}.$$

By the convolution formula for densities, the righthand side is $(2\pi)^k$ times the density $p_{X+\sigma Z}(y)$ of the sum of X and σZ for a standard normal vector Z that is independent of X. Conclude that if X and Y have the same characteristic function, then the vectors $X + \sigma Z$ and $Y + \sigma Z$ have the same density and hence are equal in distribution for every $\sigma > 0$. By Slutsky's lemma $X + \sigma Z \rightsquigarrow X$ as $\sigma \downarrow 0$, and similarly for Y. Thus X and Y are equal in distribution. ■

The characteristic function of a sum of independent variables equals the product of the characteristic functions of the individual variables. This observation, combined with Lévy's theorem, yields simple proofs of both the law of large numbers and the central limit theorem.

2.16 ***Proposition (Weak law of large numbers).*** *Let $Y_1, \dots, Y_n$ be i.i.d. random variables with characteristic function ϕ. Then $\overline{Y}_n \xrightarrow{\mathrm{P}} \mu$ for a real number μ if and only if ϕ is differentiable at zero with $i\mu = \phi'(0)$.*

Proof. We only prove that differentiability is sufficient. For the converse, see, for example, [127, p. 52]. Because $\phi(0) = 1$, differentiability of ϕ at zero means that $\phi(t) = 1 + t\phi'(0) + o(t)$ as $t \to 0$. Thus, by Fubini's theorem, for each fixed t and $n \to \infty$,

$$\mathrm{E}e^{it\overline{Y}_n} = \phi^n\left(\frac{t}{n}\right) = \left(1 + \frac{t}{n}i\mu + o\left(\frac{1}{n}\right)\right)^n \to e^{it\mu}.$$

The right side is the characteristic function of the constant variable μ. By Lévy's theorem, $\overline{Y}_n$ converges in distribution to μ. Convergence in distribution to a constant is the same as convergence in probability. ■

A sufficient but not necessary condition for $\phi(t) = \mathrm{E}e^{itY}$ to be differentiable at zero is that $\mathrm{E}|Y| < \infty$. In that case the dominated convergence theorem allows differentiation under the expectation sign, and we obtain

$$\phi'(t) = \frac{d}{dt}\mathrm{E}e^{itY} = \mathrm{E}iYe^{itY}.$$

In particular, the derivative at zero is $\phi'(0) = i\mathrm{E}Y$ and hence $\overline{Y}_n \xrightarrow{\mathrm{P}} \mathrm{E}Y_1$.

If $\mathrm{E}Y^2 < \infty$, then the Taylor expansion can be carried a step further and we can obtain a version of the central limit theorem.

2.17 ***Proposition (Central limit theorem).*** *Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $\mathrm{E}Y_i = 0$ and $\mathrm{E}Y_i^2 = 1$. Then the sequence $\sqrt{n}\,\overline{Y}_n$ converges in distribution to the standard normal distribution.*

Proof. A second differentiation under the expectation sign shows that $\phi''(0) = i^2\mathrm{E}Y^2$. Because $\phi'(0) = i\mathrm{E}Y = 0$, we obtain

$$\mathrm{E}e^{it\sqrt{n}\overline{Y}_n} = \phi^n\left(\frac{t}{\sqrt{n}}\right) = \left(1 - \frac{1}{2}\frac{t^2}{n}\mathrm{E}Y^2 + o\left(\frac{1}{n}\right)\right)^n \to e^{-\frac{1}{2}t^2\mathrm{E}Y^2}.$$

The right side is the characteristic function of the normal distribution with mean zero and variance $\mathrm{E}Y^2$. The proposition follows from Lévy's continuity theorem. ■

The characteristic function $t \mapsto \mathrm{E}e^{it^TX}$ of a vector X is determined by the set of all characteristic functions $u \mapsto \mathrm{E}e^{iu(t^TX)}$ of linear combinations t^TX of the components of X. Therefore, Lévy's continuity theorem implies that weak convergence of vectors is equivalent to weak convergence of linear combinations:

$$X_n \rightsquigarrow X \quad \text{if and only if} \quad t^TX_n \rightsquigarrow t^TX \quad \text{for all} \quad t \in \mathbb{R}^k.$$

This is known as the *Cramér-Wold device*. It allows to reduce higher-dimensional problems to the one-dimensional case.

2.18 ***Example (Multivariate central limit theorem).*** Let $Y_1, Y_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^k$ with mean vector $\mu = \mathrm{E}Y_1$ and covariance matrix $\Sigma = \mathrm{E}(Y_1 - \mu)(Y_1 - \mu)^T$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \mu) = \sqrt{n}(\overline{Y}_n - \mu) \rightsquigarrow N_k(0, \Sigma).$$

(The sum is taken coordinatewise.) By the Cramér-Wold device, this can be proved by finding the limit distribution of the sequences of real variables

$$t^T\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \mu)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (t^TY_i - t^T\mu).$$

Because the random variables $t^TY_1 - t^T\mu, t^TY_2 - t^T\mu, \ldots$ are i.i.d. with zero mean and variance $t^T\Sigma t$, this sequence is asymptotically $N_1(0, t^T\Sigma t)$-distributed by the univariate central limit theorem. This is exactly the distribution of t^TX if X possesses an $N_k(0, \Sigma)$ distribution. □

*2.4 Almost-Sure Representations

Convergence in distribution certainly does not imply convergence in probability or almost surely. However, the following theorem shows that a given sequence $X_n \rightsquigarrow X$ can always be replaced by a sequence $\tilde{X}_n \rightsquigarrow \tilde{X}$ that is, marginally, equal in distribution and converges almost surely. This construction is sometimes useful and has been put to good use by some authors, but we do not use it in this book.

2.19 ***Theorem (Almost-sure representations).*** *Suppose that the sequence of random vectors X_n converges in distribution to a random vector X_0. Then there exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{U}}, \tilde{P})$ and random vectors $\tilde{X}_n$ defined on it such that $\tilde{X}_n$ is equal in distribution to X_n for every $n \geq 0$ and $\tilde{X}_n \to \tilde{X}_0$ almost surely.*

Proof. For random variables we can simply define $\tilde{X}_n = F_n^{-1}(U)$ for F_n the distribution function of X_n and U an arbitrary random variable with the uniform distribution on $[0, 1]$. (The "quantile transformation," see Section 21.1.) The simplest known construction for higher-dimensional vectors is more complicated. See, for example, Theorem 1.10.4 in [146], or [41]. ■

*2.5 Convergence of Moments

By the portmanteau lemma, weak convergence $X_n \rightsquigarrow X$ implies that $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for every continuous, bounded function f. The condition that f be bounded is not superfluous: It is not difficult to find examples of a sequence $X_n \rightsquigarrow X$ and an unbounded, continuous function f for which the convergence fails. In particular, in general convergence in distribution does not imply convergence $\mathrm{E}X_n^p \to \mathrm{E}X^p$ of moments. However, in many situations such convergence occurs, but it requires more effort to prove it.

A sequence of random variables Y_n is called *asymptotically uniformly integrable* if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{E}|Y_n| 1\{|Y_n| > M\} = 0.$$

Uniform integrability is the missing link between convergence in distribution and convergence of moments.

2.20 ***Theorem.*** *Let $f : \mathbb{R}^k \mapsto \mathbb{R}$ be measurable and continuous at every point in a set C. Let $X_n \rightsquigarrow X$ where X takes its values in C. Then $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ if and only if the sequence of random variables $f(X_n)$ is asymptotically uniformly integrable.*

Proof. We give the proof only in the most interesting direction. (See, for example, [146] (p. 69) for the other direction.) Suppose that $Y_n = f(X_n)$ is asymptotically uniformly integrable. Then we show that $\mathrm{E}Y_n \to \mathrm{E}Y$ for $Y = f(X)$. Assume without loss of generality that Y_n is nonnegative; otherwise argue the positive and negative parts separately. By the continuous mapping theorem, $Y_n \rightsquigarrow Y$. By the triangle inequality,

$$|\mathrm{E}Y_n - \mathrm{E}Y| \leq |\mathrm{E}Y_n - \mathrm{E}Y_n \wedge M| + |\mathrm{E}Y_n \wedge M - \mathrm{E}Y \wedge M| + |\mathrm{E}Y \wedge M - \mathrm{E}Y|.$$

Because the function $y \mapsto y \wedge M$ is continuous and bounded on $[0, \infty)$, it follows that the middle term on the right converges to zero as $n \to \infty$. The first term is bounded above by

$\mathrm{E}Y_n 1\{Y_n > M\}$, and converges to zero as $n \to \infty$ followed by $M \to \infty$, by the uniform integrability. By the portmanteau lemma (iv), the third term is bounded by the lim inf as $n \to \infty$ of the first and hence converges to zero as $M \uparrow \infty$. ∎

2.21 ***Example.*** Suppose X_n is a sequence of random variables such that $X_n \rightsquigarrow X$ and $\limsup \mathrm{E}|X_n|^p < \infty$ for some p. Then all moments of order strictly less than p converge also: $\mathrm{E}X_n^k \to \mathrm{E}X^k$ for every $k < p$.

By the preceding theorem, it suffices to prove that the sequence X_n^k is asymptotically uniformly integrable. By Markov's inequality

$$\mathrm{E}|X_n|^k 1\left\{|X_n|^k \geq M\right\} \leq M^{1-p/k}\,\mathrm{E}|X_n|^p.$$

The limit superior, as $n \to \infty$ followed by $M \to \infty$, of the right side is zero if $k < p$. □

The moment function $p \mapsto \mathrm{E}X^p$ can be considered a transform of probability distributions, just as can the characteristic function. In general, it is not a true transform in that it does determine a distribution uniquely only under additional assumptions. If a limit distribution is uniquely determined by its moments, this transform can still be used to establish weak convergence.

2.22 ***Theorem.*** *Let X_n and X be random variables such that $\mathrm{E}X_n^p \to \mathrm{E}X^p < \infty$ for every $p \in \mathbb{N}$. If the distribution of X is uniquely determined by its moments, then $X_n \rightsquigarrow X$.*

Proof. Because $\mathrm{E}X_n^2 = O(1)$, the sequence X_n is uniformly tight, by Markov's inequality. By Prohorov's theorem, each subsequence has a further subsequence that converges weakly to a limit Y. By the preceding example the moments of Y are the limits of the moments of the subsequence. Thus the moments of Y are identical to the moments of X. Because, by assumption, there is only one distribution with this set of moments, X and Y are equal in distribution. Conclude that every subsequence of X_n has a further subsequence that converges in distribution to X. This implies that the whole sequence converges to X. ∎

2.23 ***Example.*** The normal distribution is uniquely determined by its moments. (See, for example, [123] or [133, p. 293].) Thus $\mathrm{E}X_n^p \to 0$ for odd p and $\mathrm{E}X_n^p \to (p-1)(p-3)\cdots 1$ for even p implies that $X_n \rightsquigarrow N(0, 1)$. The converse is false. □

*2.6 Convergence-Determining Classes

A class $\mathcal{F}$ of functions $f : \mathbb{R}^k \to \mathbb{R}$ is called *convergence-determining* if for every sequence of random vectors X_n the convergence $X_n \rightsquigarrow X$ is equivalent to $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for every $f \in \mathcal{F}$. By definition the set of all bounded continuous functions is convergence-determining, but so is the smaller set of all differentiable functions, and many other classes. The set of all indicator functions $1_{(-\infty,t]}$ would be convergence-determining if we would restrict the definition to limits X with continuous distribution functions. We shall have occasion to use the following results. (For proofs see Corollary 1.4.5 and Theorem 1.12.2, for example, in [146].)

2.24 ***Lemma.*** *On $\mathbb{R}^k = \mathbb{R}^l \times \mathbb{R}^m$ the set of functions $(x, y) \mapsto f(x)g(y)$ with f and g ranging over all bounded, continuous functions on $\mathbb{R}^l$ and $\mathbb{R}^m$, respectively, is convergence-determining.*

2.25 ***Lemma.*** *There exists a countable set of continuous functions $f : \mathbb{R}^k \mapsto [0, 1]$ that is convergence-determining and, moreover, $X_n \rightsquigarrow X$ implies that $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ uniformly in $f \in \mathcal{F}$.*

*2.7 Law of the Iterated Logarithm

The law of the iterated logarithm is an intriguing result but appears to be of less interest to statisticians. It can be viewed as a refinement of the strong law of large numbers. If $Y_1, Y_2, \ldots$ are i.i.d. random variables with mean zero, then $Y_1 + \cdots + Y_n = o(n)$ almost surely by the strong law. The law of the iterated logarithm improves this order to $O(\sqrt{n \log \log n})$, and even gives the proportionality constant.

2.26 ***Proposition (Law of the iterated logarithm).*** *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with mean zero and variance 1. Then*

$$\limsup_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{\sqrt{n \log \log n}} = \sqrt{2}, \qquad a.s.$$

Conversely, if this statement holds for both Y_i and $-Y_i$, then the variables have mean zero and variance 1.

The law of the iterated logarithm gives an interesting illustration of the difference between almost sure and distributional statements. Under the conditions of the proposition, the sequence $n^{-1/2}(Y_1 + \cdots + Y_n)$ is asymptotically normally distributed by the central limit theorem. The limiting normal distribution is spread out over the whole real line. Apparently division by the factor $\sqrt{\log \log n}$ is exactly right to keep $n^{-1/2}(Y_1 + \cdots + Y_n)$ within a compact interval, eventually.

A simple application of Slutsky's lemma gives

$$Z_n := \frac{Y_1 + \cdots + Y_n}{\sqrt{n \log \log n}} \xrightarrow{\mathrm{P}} 0.$$

Thus Z_n is with high probability contained in the interval $(-\varepsilon, \varepsilon)$ eventually, for any $\varepsilon > 0$. This appears to contradict the law of the iterated logarithm, which asserts that Z_n reaches the interval $(\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon)$ infinitely often with probability one. The explanation is that the set of ω such that $Z_n(\omega)$ is in $(-\varepsilon, \varepsilon)$ or $(\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon)$ fluctuates with n. The convergence in probability shows that at any advanced time a very large fraction of ω have $Z_n(\omega) \in (-\varepsilon, \varepsilon)$. The law of the iterated logarithm shows that for each particular ω the sequence $Z_n(\omega)$ drops in and out of the interval $(\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon)$ infinitely often (and hence out of $(-\varepsilon, \varepsilon)$).

The implications for statistics can be illustrated by considering confidence statements. If μ and 1 are the true mean and variance of the sample $Y_1, Y_2, \ldots$, then the probability that

$$\overline{Y}_n - \frac{2}{\sqrt{n}} \le \mu \le \overline{Y}_n + \frac{2}{\sqrt{n}}$$

converges to $\Phi(2) - \Phi(-2) \approx 95\%$. Thus the given interval is an asymptotic confidence interval of level approximately 95%. (The confidence level is exactly $\Phi(2) - \Phi(-2)$ if the observations are normally distributed. This may be assumed in the following; the accuracy of the approximation is not an issue in this discussion.) The point $\mu = 0$ is contained in the interval if and only if the variable Z_n satisfies

$$|Z_n| \leq \frac{2}{\sqrt{\log\log n}}.$$

Assume that $\mu = 0$ is the true value of the mean, and consider the following argument. By the law of the iterated logarithm, we can be sure that Z_n hits the interval $(\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon)$ infinitely often. The expression $2/\sqrt{\log\log n}$ is close to zero for large n. Thus we can be sure that the true value $\mu = 0$ is outside the confidence interval infinitely often.

How can we solve the paradox that the usual confidence interval is wrong infinitely often? There appears to be a conceptual problem if it is imagined that a statistician collects data in a sequential manner, computing a confidence interval for every n. However, although the frequentist interpretation of a confidence interval is open to the usual criticism, the paradox does not seem to rise within the frequentist framework. In fact, from a frequentist point of view the curious conclusion is reasonable. Imagine 100 statisticians, all of whom set 95% confidence intervals in the usual manner. They all receive one observation per day and update their confidence intervals daily. Then every day about five of them should have a false interval. It is only fair that as the days go by all of them take turns in being unlucky, and that the same five do not have it wrong all the time. This, indeed, happens according to the law of the iterated logarithm.

The paradox may be partly caused by the feeling that with a growing number of observations, the confidence intervals should become better. In contrast, the usual approach leads to errors with certainty. However, this is only true if the usual approach is applied naively in a sequential set-up. In practice one would do a genuine sequential analysis (including the use of a stopping rule) or change the confidence level with n.

There is also another reason that the law of the iterated logarithm is of little practical consequence. The argument in the preceding paragraphs is based on the assumption that $2/\sqrt{\log\log n}$ is close to zero and is nonsensical if this quantity is larger than $\sqrt{2}$. Thus the argument requires at least $n \geq 1619$, a respectable number of observations.

*2.8 Lindeberg-Feller Theorem

Central limit theorems are theorems concerning convergence in distribution of sums of random variables. There are versions for dependent observations and nonnormal limit distributions. The Lindeberg-Feller theorem is the simplest extension of the classical central limit theorem and is applicable to independent observations with finite variances.

2.27 ***Proposition (Lindeberg-Feller central limit theorem).*** *For each n let* $Y_{n,1}, \ldots, Y_{n,k_n}$ *be independent random vectors with finite variances such that*

$$\sum_{i=1}^{k_n} \mathrm{E}\|Y_{n,i}\|^2 1\{\|Y_{n,i}\| > \varepsilon\} \to 0, \qquad \textit{every } \varepsilon > 0,$$

$$\sum_{i=1}^{k_n} \mathrm{Cov}\, Y_{n,i} \to \Sigma.$$

Then the sequence $\sum_{i=1}^{k_n}(Y_{n,i} - \mathrm{E}Y_{n,i})$ *converges in distribution to a normal* $N(0, \Sigma)$ *distribution.*

A result of this type is necessary to treat the asymptotics of, for instance, regression problems with fixed covariates. We illustrate this by the linear regression model. The application is straightforward but notationally a bit involved. Therefore, at other places in the manuscript we find it more convenient to assume that the covariates are a random sample, so that the ordinary central limit theorem applies.

2.28 ***Example (Linear regression).*** In the linear regression problem, we observe a vector $Y = X\beta + e$ for a known $(n \times p)$ matrix X of full rank, and an (unobserved) error vector e with i.i.d. components with mean zero and variance σ^2. The least squares estimator of β is

$$\hat\beta = (X^T X)^{-1} X^T Y.$$

This estimator is unbiased and has covariance matrix $\sigma^2 (X^T X)^{-1}$. If the error vector e is normally distributed, then $\hat\beta$ is exactly normally distributed. Under reasonable conditions on the design matrix, the least squares estimator is asymptotically normally distributed for a large range of error distributions. Here we fix p and let n tend to infinity.

This follows from the representation

$$(X^T X)^{1/2}(\hat\beta - \beta) = (X^T X)^{-1/2} X^T e = \sum_{i=1}^n a_{ni} e_i,$$

where $a_{n1}, \ldots, a_{nn}$ are the columns of the $(p \times n)$ matrix $(X^T X)^{-1/2} X^T =: A$. This sequence is asymptotically normal if the vectors $a_{n1}e_1, \ldots, a_{nn}e_n$ satisfy the Lindeberg conditions. The norming matrix $(X^T X)^{1/2}$ has been chosen to ensure that the vectors in the display have covariance matrix $\sigma^2 I$ for every n. The remaining condition is

$$\sum_{i=1}^n \|a_{ni}\|^2 \mathrm{E}e_i^2 1\big\{\|a_{ni}\||e_i| > \varepsilon\big\} \to 0.$$

This can be simplified to other conditions in several ways. Because $\sum \|a_{ni}\|^2 = \operatorname{trace}(AA^T) = p$, it suffices that $\max \mathrm{E}e_i^2 1\big\{\|a_{ni}\||e_i| > \varepsilon\big\} \to 0$, which is equivalent to

$$\max_{1 \le i \le n} \|a_{ni}\| \to 0.$$

Alternatively, the expectation $\mathrm{E}e^2 1\big\{a|e| > \varepsilon\big\}$ can be bounded by $\varepsilon^{-k}\mathrm{E}|e|^{k+2}a^k$ and a second set of sufficient conditions is

$$\sum_{i=1}^n \|a_{ni}\|^k \to 0; \qquad \mathrm{E}|e_1|^k < \infty, \qquad (k > 2).$$

Both sets of conditions are reasonable. Consider for instance the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$. Then

$$(X^T X)^{-1/2} X^T = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix}^{-1/2} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}.$$

It is reasonable to assume that the sequences $\overline{x}$ and $\overline{x^2}$ are bounded. Then the first matrix on the right behaves like a fixed matrix, and the conditions for asymptotic normality simplify to

$$\max_{1\le i\le n} |x_i| = o(n^{1/2}); \quad \text{or} \quad n^{1-k/2}\overline{|x|^k} \to 0, \qquad \mathrm{E}|e_1|^k < \infty.$$

Every reasonable design satisfies these conditions. □

*2.9 Convergence in Total Variation

A sequence of random variables converges in *total variation* to a variable X if

$$\sup_B \big|\mathrm{P}(X_n \in B) - \mathrm{P}(X \in B)\big| \to 0,$$

where the supremum is taken over all measurable sets B. In view of the portmanteau lemma, this type of convergence is stronger than convergence in distribution. Not only is it required that the sequence $\mathrm{P}(X_n \in B)$ converges for every Borel set B, the convergence must also be uniform in B. Such strong convergence occurs less frequently and is often more than necessary, whence the concept is less useful.

A simple sufficient condition for convergence in total variation is pointwise convergence of densities. If X_n and X have densities p_n and p with respect to a measure μ, then

$$\sup_B \big|\mathrm{P}(X_n \in B) - \mathrm{P}(X \in B)\big| = \frac{1}{2}\int |p_n - p|\,d\mu.$$

Thus, convergence in total variation can be established by convergence theorems for integrals from measure theory. The following proposition, which should be compared with the monotone and dominated convergence theorems, is most appropriate.

2.29 ***Proposition.*** *Suppose that f_n and f are arbitrary measurable functions such that $f_n \to f$ μ-almost everywhere (or in μ-measure) and* $\limsup \int |f_n|^p\,d\mu \le \int |f|^p\,d\mu < \infty$, *for some $p \ge 1$ and measure μ. Then* $\int |f_n - f|^p\,d\mu \to 0$.

Proof. By the inequality $(a+b)^p \le 2^p a^p + 2^p b^p$, valid for every $a, b \ge 0$, and the assumption, $0 \le 2^p|f_n|^p + 2^p|f|^p - |f_n - f|^p \to 2^{p+1}|f|^p$ almost everywhere. By Fatou's lemma,

$$\begin{aligned}\int 2^{p+1}|f|^p\,d\mu &\le \liminf \int \big(2^p|f_n|^p + 2^p|f|^p - |f_n - f|^p\big)\,d\mu \\ &\le 2^{p+1}\int |f|^p\,d\mu - \limsup \int |f_n - f|^p\,d\mu,\end{aligned}$$

by assumption. The proposition follows. ■

2.30 ***Corollary (Scheffé).*** *Let X_n and X be random vectors with densities p_n and p with respect to a measure μ. If $p_n \to p$ μ-almost everywhere, then the sequence X_n converges to X in total variation.*

The central limit theorem is usually formulated in terms of convergence in distribution. Often it is valid in terms of the total variation distance, in the sense that

$$\sup_B \left| \mathrm{P}(Y_1 + \cdots + Y_n \in B) - \int_B \frac{1}{\sqrt{n}\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-n\mu)^2/n\sigma^2}\, dx \right| \to 0.$$

Here μ and σ^2 are mean and variance of the Y_i, and the supremum is taken over all Borel sets. An integrable characteristic function, in addition to a finite second moment, suffices.

2.31 ***Theorem (Central limit theorem in total variation).*** *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with finite second moment and characteristic function ϕ such that $\int |\phi(t)|^\nu\, dt < \infty$ for some $\nu \geq 1$. Then $Y_1 + \cdots + Y_n$ satisfies the central limit theorem in total variation.*

Proof. It can be assumed without loss of generality that $\mathrm{E}Y_1 = 0$ and $\mathrm{var}\, Y_1 = 1$. By the inversion formula for characteristic functions (see [47, p. 509]), the density p_n of $Y_1 + \cdots + Y_n/\sqrt{n}$ can be written

$$p_n(x) = \frac{1}{2\pi} \int e^{-itx} \phi\left(\frac{t}{\sqrt{n}}\right)^n dt.$$

By the central limit theorem and Lévy's continuity theorem, the integrand converges to $e^{-itx} \exp(-\frac{1}{2}t^2)$. It will be shown that the integral converges to

$$\frac{1}{2\pi} \int e^{-itx} e^{-\frac{1}{2}t^2}\, dt = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}.$$

Then an application of Scheffé's theorem concludes the proof.

The integral can be split into two parts. First, for every $\varepsilon > 0$,

$$\int_{|t|>\varepsilon\sqrt{n}} \left| e^{-itx} \phi\left(\frac{t}{\sqrt{n}}\right)^n \right| dt \leq \sqrt{n} \sup_{|t|>\varepsilon} |\phi(t)|^{n-\nu} \int |\phi(t)|^\nu\, dt.$$

Here $\sup_{|t|>\varepsilon} |\phi(t)| < 1$ by the Riemann-Lebesgue lemma and because ϕ is the characteristic function of a nonlattice distribution (e.g., [47, pp. 501, 513]). Thus, the first part of the integral converges to zero geometrically fast.

Second, a Taylor expansion yields that $\phi(t) = 1 - \frac{1}{2}t^2 + o(t^2)$ as $t \to 0$, so that there exists $\varepsilon > 0$ such that $|\phi(t)| \leq 1 - t^2/4$ for every $|t| < \varepsilon$. It follows that

$$\left| e^{-itx} \phi\left(\frac{t}{\sqrt{n}}\right)^n \right| 1\{|t| \leq \varepsilon\sqrt{n}\} \leq \left(1 - \frac{t^2}{4n}\right)^n \leq e^{-t^2/4}.$$

The proof can be concluded by applying the dominated convergence theorem to the remaining part of the integral. ∎

Notes

The results of this chapter can be found in many introductions to probability theory. A standard reference for weak convergence theory is the first chapter of [11]. Another very readable introduction is [41]. The theory of this chapter is extended to random elements with values in general metric spaces in Chapter 18.

PROBLEMS

1. If X_n possesses a t-distribution with n degrees of freedom, then $X_n \rightsquigarrow N(0,1)$ as $n \to \infty$. Show this.
2. Does it follow immediately from the result of the previous exercise that $\mathrm{E}X_n^p \to \mathrm{E}N(0,1)^p$ for every $p \in \mathbb{N}$? Is this true?
3. If $X_n \rightsquigarrow N(0,1)$ and $Y_n \xrightarrow{\mathrm{P}} \sigma$, then $X_n Y_n \rightsquigarrow N(0, \sigma^2)$. Show this.
4. In what sense is a chi-square distribution with n degrees of freedom approximately a normal distribution?
5. Find an example of sequences such that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$, but the joint sequence (X_n, Y_n) does not converge in law.
6. If X_n and Y_n are independent random vectors for every n, then $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ imply that $(X_n, Y_n) \rightsquigarrow (X, Y)$, where X and Y are independent. Show this.
7. If every X_n and X possess discrete distributions supported on the integers, then $X_n \rightsquigarrow X$ if and only if $\mathrm{P}(X_n = x) \to \mathrm{P}(X = x)$ for every integer x. Show this.
8. If $\mathrm{P}(X_n = i/n) = 1/n$ for every $i = 1, 2, \ldots, n$, then $X_n \rightsquigarrow X$, but there exist Borel sets with $\mathrm{P}(X_n \in B) = 1$ for every n, but $\mathrm{P}(X \in B) = 0$. Show this.
9. If $\mathrm{P}(X_n = x_n) = 1$ for numbers x_n and $x_n \to x$, then $X_n \rightsquigarrow x$. Prove this
   (i) by considering distributions functions
   (ii) by using Theorem 2.7.
10. State the rule $o_P(1) + O_P(1) = O_P(1)$ in terms of random vectors and show its validity.
11. In what sense is it true that $o_P(1) = O_P(1)$? Is it true that $O_P(1) = o_P(1)$?
12. The rules given by Lemma 2.12 are not simple plug-in rules.
    (i) Give an example of a function R with $R(h) = o(\|h\|)$ as $h \to 0$ and a sequence of random variables X_n such that $R(X_n)$ is not equal to $o_P(X_n)$.
    (ii) Given an example of a function R such $R(h) = O(\|h\|)$ as $h \to 0$ and a sequence of random variables X_n such that $X_n = O_P(1)$ but $R(X_n)$ is not equal to $O_P(X_n)$.
13. Find an example of a sequence of random variables such that $X_n \rightsquigarrow 0$, but $\mathrm{E}X_n \to \infty$.
14. Find an example of a sequence of random variables such that $X_n \xrightarrow{\mathrm{P}} 0$, but X_n does not converge almost surely.
15. Let $X_1, \ldots, X_n$ be i.i.d. with density $f_{\lambda,a}(x) = \lambda e^{-\lambda(x-a)} 1\{x \geq a\}$. Calculate the maximum likelihood estimator of $(\hat{\lambda}_n, \hat{a}_n)$ of (λ, a) and show that $(\hat{\lambda}_n, \hat{a}_n) \xrightarrow{\mathrm{P}} (\lambda, a)$.
16. Let $X_1, \ldots, X_n$ be i.i.d. standard normal variables. Show that the vector $U = (X_1, \ldots, X_n)/N$, where $N^2 = \sum_{i=1}^n X_i^2$, is uniformly distributed over the unit sphere S^{n-1} in $\mathbb{R}^n$, in the sense that U and OU are identically distributed for every orthogonal transformation O of $\mathbb{R}^n$.
17. For each n, let U_n be uniformly distributed over the unit sphere S^{n-1} in $\mathbb{R}^n$. Show that the vectors $\sqrt{n}(U_{n,1}, U_{n,2})$ converge in distribution to a pair of independent standard normal variables.
18. If $\sqrt{n}(T_n - \theta)$ converges in distribution, then T_n converges in probability to θ. Show this.
19. If $\mathrm{E}X_n \to \mu$ and $\operatorname{var} X_n \to 0$, then $X_n \xrightarrow{\mathrm{P}} \mu$. Show this.
20. If $\sum_{n=1}^{\infty} \mathrm{P}(|X_n| > \varepsilon) < \infty$ for every $\varepsilon > 0$, then X_n converges almost surely to zero. Show this.
21. Use characteristic functions to show that binomial$(n, \lambda/n) \rightsquigarrow$ Poisson(λ). Why does the central limit theorem not hold?
22. If $X_1, \ldots, X_n$ are i.i.d. standard Cauchy, then $\overline{X}_n$ is standard Cauchy.
    (i) Show this by using characteristic functions
    (ii) Why does the weak law not hold?
23. Let $X_1, \ldots, X_n$ be i.i.d. with finite fourth moment. Find constants a, b, and c_n such that the sequence $c_n(\overline{X}_n - a, \overline{X_n^2} - b)$ converges in distribution, and determine the limit law. Here $\overline{X}_n$ and $\overline{X_n^2}$ are the averages of the X_i and the X_i^2, respectively.