# Chapter 2

# Weak Convergence

Chapter 1 discussed limits of sequences of constants, either scalar-valued or vector-valued. Chapters 2 and 3 extend this notion by defining what it means for a sequence of random variables to have a limit. As it turns out, there is more than one sensible way to do this.

Chapters 2 and 4 (and, to a lesser extent, Chapter 3) lay the theoretical groundwork for nearly all of the statistical topics that will follow. While the material in Chapter 2 is essential, readers may wish to skip Chapter 3 on a first reading. Furthermore, as is common throughout the book, some of the proofs here have been relegated to the exercises, meaning that readers may study these proofs as much or as little as they desire.

## 2.1 Modes of Convergence

Whereas the limit of a constant sequence is unequivocally expressed by Definition 1.30, in the case of random variables there are several ways to define the convergence of a sequence. This section discusses three such definitions, or modes, of convergence; Section 3.1 presents a fourth.

### 2.1.1 Convergence in Probability

What does it mean for the sequence $X_1, X_2, \ldots$ of random variables to converge to, say, the random variable $X$? In what case should one write $X_n \to X$? We begin by considering a definition of convergence that requires that $X_n$ and $X$ be defined on the same sample space. For this form of convergence, called convergence in probability, the absolute difference $|X_n - X|$, itself a random variable, should be arbitrarily close to zero with probability arbitrarily close to one. More precisely, we make the following definition.

**Definition 2.1** Let $X_n$ and $X$ be defined on the same probability space. We say that $X_n$ converges in probability to $X$, written $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$P(|X_n - X| < \epsilon) \to 1 \text{ as } n \to \infty. \tag{2.1}$$

It is very common that the $X$ in Definition 2.1 is a constant, say $X \equiv c$. In such cases, we simply write $X_n \xrightarrow{P} c$. There is no harm in replacing $X$ by $c$ in Definition 2.1 because any constant may be defined as a random variable on any sample space. In the most common statistical usage of convergence to a constant, the random variable $X_n$ is some estimator of a particular parameter, say $g(\theta)$:

**Definition 2.2** If $X_n \xrightarrow{P} g(\theta)$, $X_n$ is said to be **consistent** (or weakly consistent) for $g(\theta)$.

As the name suggests, weak consistency is implied by a concept called "strong consistency," a topic to be explored in Chapter 3. Note that "consistency," used without the word "strong" or "weak," generally refers to weak consistency.

**Example 2.3** Suppose that $Y_1, Y_2, \ldots$ are independent and identically distributed uniform $(0, \theta)$ random variables, where $\theta$ is an unknown positive constant. For $n \geq 1$, let $X_n$ be defined as the largest value among $Y_1$ through $Y_n$: That is, $X_n \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} X_i$. Then we may show that $X_n$ is a consistent estimator of $\theta$ as follows:

By Definition 2.1, we wish to show that for an arbitrary $\epsilon > 0$, $P(|X_n - \theta| < \epsilon) \to 1$ as $n \to \infty$. In this particular case, we can evaluate $P(|X_n - \theta| < \epsilon)$ directly by noting that $X_n$ cannot possibly be larger than $\theta$, so that

$$P(|X_n - \theta| < \epsilon) = P(X_n > \theta - \epsilon) = 1 - P(X_n \leq \theta - \epsilon).$$

Now the maximum $X_n$ is less than a constant if and only if each of the random variables $Y_1, \ldots, Y_n$ is less than that constant. Therefore, by independence,

$$P(X_n \leq \theta - \epsilon) = [P(Y_1 \leq \theta - \epsilon)]^n = \left[1 - \frac{\epsilon}{\theta}\right]^n$$

as long as $\epsilon < \theta$. (If $\epsilon \geq \theta$, the above probability is simply zero.) Since any constant $c$ in the open interval $(0, 1)$ has the property that $c^n \to 0$ as $n \to \infty$, we conclude that $P(X_n \leq \theta - \epsilon) \to 0$ as desired. ■

There are probabilistic analogues of the $o$ and $O$ notations of Section 1.3 that are frequently used in the literature but that apply to random variable sequences instead of constant sequences.

**Definition 2.4** We write $X_n = o_P(Y_n)$ if $X_n/Y_n \xrightarrow{P} 0$.

**Definition 2.5** We write $X_n = O_P(Y_n)$ if for every $\epsilon > 0$, there exist $M$ and $N$ such that

$$P\left(\left|\frac{X_n}{Y_n}\right| < M\right) > 1 - \epsilon \text{ for all } n > N.$$

In particular, it is common to write $o_P(1)$, to denote a sequence of random variables that converges to zero in probability, or $X_n = O_P(1)$, to state that $X_n$ is bounded in probability (see Exercise 2.7 for a definition of bounded in probability).

**Example 2.6** In Example 2.3, we showed that if $Y_1, Y_2, \ldots$ are independent and identically distributed uniform $(0, \theta)$ random variables, then

$$\max_{1 \le i \le n} Y_i \xrightarrow{P} \theta \qquad \text{as } n \to \infty.$$

Equivalently, we may say that

$$\max_{1 \le i \le n} Y_i = \theta + o_P(1) \qquad \text{as } n \to \infty. \tag{2.2}$$

It is also correct to write

$$\max_{1 \le i \le n} Y_i = \theta + O_P(1) \qquad \text{as } n \to \infty, \tag{2.3}$$

though statement (2.3) is less informative than statement (2.2). On the other hand, we will see in Example 6.1 that statement (2.3) may be sharpened considerably — and made even more informative than statement (2.2) — by writing

$$\max_{1 \le i \le n} Y_i = \theta + O_P\left(\frac{1}{n}\right) \qquad \text{as } n \to \infty.$$

### 2.1.2 Convergence in Distribution

As the name suggests, convergence in distribution has to do with convergence of the distribution functions of random variables. Given a random variable $X$, the distribution function of $X$ is the function

$$F(x) = P(X \le x). \tag{2.4}$$

Any distribution function $F(x)$ is nondecreasing and right-continuous, and it has limits $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. Conversely, any function $F(x)$ with these properties is a distribution function for some random variable.

It is not enough to define convergence in distribution as simple pointwise convergence of a sequence of distribution functions; there are technical reasons that such a simplistic definition fails to capture any useful concept of convergence of random variables. These reasons are illustrated by the following two examples.

**Example 2.7** Let $X_n$ be normally distributed with mean 0 and variance $n$. Then the distribution function of $X_n$ is $\Phi(x/\sqrt{n})$, where $\Phi(z)$ denotes the standard normal distribution function. Because $\Phi(0) = 1/2$, we see that for any fixed point $x$, the distribution function of $X_n$ converges to $1/2$ at that point as $n \to \infty$. But the function that is constant at $1/2$ is not a distribution function. Therefore, this example shows that not all convergent sequences of distribution functions have limits that are distribution functions.

**Example 2.8** By any sensible definition of convergence, $1/n$ should converge to 0. But consider the distribution functions $F_n(x) = I\{x \ge 1/n\}$ and $F(x) = I\{x \ge 0\}$ corresponding to the constant random variables $1/n$ and 0. We do *not* have pointwise convergence of $F_n(x)$ to $F(x)$, since $F_n(0) = 0$ for all $n$ but $F(0) = 1$. Note, however, that $F_n(x) \to F(x)$ is true for all $x \ne 0$. Not coincidentally, the point $x = 0$ is the only point at which the function $F(x)$ is not continuous.

To write a sensible definition of convergence in distribution, Example 2.7 demonstrates that we should require that the limit of distribution functions be a distribution function, while Example 2.8 indicates that we should exclude points where $F(x)$ is not continuous. We therefore arrive at the following definition:

**Definition 2.9** Suppose that $X$ has distribution function $F(x)$ and that $X_n$ has distribution function $F_n(x)$ for each $n$. Then we say $X_n$ converges in distribution to $X$, written $X_n \stackrel{d}{\to} X$, if $F_n(x) \to F(x)$ as $n \to \infty$ for all $x$ at which $F(x)$ is continuous. Convergence in distribution is sometimes called convergence in law and written $X_n \stackrel{\mathcal{L}}{\to} X$.

**Example 2.10** *The Central Limit Theorem for i.i.d. sequences:* Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) with mean $\mu$ and finite variance $\sigma^2$. Then by a result that will be covered in Chapter 4 (but which is perhaps already known to the reader),

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \stackrel{d}{\to} N(0, \sigma^2), \tag{2.5}$$

where $N(0, \sigma^2)$ denotes a random variable with mean 0 and variance $\sigma^2$. Note that the expressions on either side of the $\stackrel{d}{\to}$ symbol may sometimes be random variables, sometimes distribution functions or other notations indicating certain

> distributions. The meaning is always clear even if the notation is sometimes inconsistent.
>
> Note that the distribution on the right hand side of (2.5) does not depend on $n$. Indeed, as a general rule, the right side of an arrow should *never* depend on $n$ when the arrow conveys the idea "as $n \to \infty$." By Definition 2.5, the right side of (2.5) is $O_P(1)$. We may therefore write (after dividing through by $\sqrt{n}$ and adding $\mu$)
>
> $$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu + O_P\left(\frac{1}{\sqrt{n}}\right) \qquad \text{as } n \to \infty,$$
>
> which is less specific than expression (2.5) but expresses at a glance the *rate* of convergence of the sample mean to $\mu$.

Could $X_n \xrightarrow{d} X$ imply $X_n \xrightarrow{P} X$? Not in general: Since convergence in distribution only involves distribution functions, $X_n \xrightarrow{d} X$ is possible even if $X_n$ and $X$ are not defined on the same sample space. However, we now prove that convergence in probability does imply convergence in distribution.

**Theorem 2.11** If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.

**Proof:** Let $F_n(x)$ and $F(x)$ denote the distribution functions of $X_n$ and $X$, respectively. Assume that $X_n \xrightarrow{P} X$. We need to show that $F_n(c) \to F(c)$, where $c$ is any point of continuity of $F(x)$.

Choose any $\epsilon > 0$. Note that whenever $X_n \le c$, it must be true that either $X \le c + \epsilon$ or $|X_n - X| > \epsilon$. This implies that

$$F_n(c) \le F(c+\epsilon) + P(|X_n - X| > \epsilon).$$

Similarly, whenever $X \le c - \epsilon$, either $X_n \le c$ or $|X_n - X| > \epsilon$, implying

$$F(c-\epsilon) \le F_n(c) + P(|X_n - X| > \epsilon).$$

We conclude that for arbitrary $n$ and $\epsilon > 0$,

$$F(c-\epsilon) - P(|X_n - X| > \epsilon) \le F_n(c) \le F(c+\epsilon) + P(|X_n - X| > \epsilon). \tag{2.6}$$

Taking both the $\liminf_n$ and the $\limsup_n$ of the above inequality, we conclude (since $X_n \xrightarrow{P} X$) that

$$F(c-\epsilon) \le \liminf_n F_n(c) \le \limsup_n F_n(c) \le F(c+\epsilon)$$

for all $\epsilon$. Since $c$ is a continuity point of $F(x)$, letting $\epsilon \to 0$ implies

$$F(c) = \liminf_n F_n(c) = \limsup_n F_n(c),$$

so we know $F_n(c) \to F(c)$ and the theorem is proved. ■

We remarked earlier that $X_n \xrightarrow{d} X$ could not possibly imply $X_n \xrightarrow{P} X$ because the latter expression requires that $X_n$ and $X$ be defined on the same sample space for every $n$. However, a constant $a$ may be considered to be a random variable defined on any sample space; thus, it is reasonable to ask whether $X_n \xrightarrow{d} a$ implies $X_n \xrightarrow{P} a$. The answer is yes:

**Theorem 2.12** $X_n \xrightarrow{d} a$ if and only if $X_n \xrightarrow{P} a$.

**Proof:** We only need to prove that $X_n \xrightarrow{d} a$ implies $X_n \xrightarrow{P} a$, since the other direction is a special case of Theorem 2.11. If $F(x)$ is the distribution function of the constant random variable $a$, then $a$ is the only point of discontinuity of $F(x)$. Therefore, for $\epsilon > 0$, $X_n \xrightarrow{d} a$ implies that $F_n(a - \epsilon) \to F(a - \epsilon) = 0$ and $F_n(a + \epsilon) \to F(a + \epsilon) = 1$ as $n \to \infty$. Therefore,

$$P(-\epsilon < X_n - a \leq \epsilon) = F_n(a + \epsilon) - F_n(a - \epsilon) \to 1,$$

which means $X_n \xrightarrow{P} a$. ■

### 2.1.3 Convergence in (quadratic) mean

**Definition 2.13** Let $k$ be a positive constant. We say that $X_n$ converges in $k$th mean to $X$, written $X_n \xrightarrow{k} X$, if

$$\mathrm{E}\, |X_n - X|^k \to 0 \text{ as } n \to \infty. \tag{2.7}$$

Two specific cases of Definition 2.13 deserve special mention. When $k = 1$, we normally omit mention of the $k$ and simply refer to the condition $\mathrm{E}\, |X_n - X| \to 0$ as *convergence in mean*. Note that convergence in mean is *not* equivalent to $\mathrm{E}\, X_n \to \mathrm{E}\, X$: For one thing, $\mathrm{E}\, X_n \to \mathrm{E}\, X$ is possible without any regard to the joint distribution of $X_n$ and $X$, whereas $\mathrm{E}\, |X_n - X| \to 0$ clearly requires that $X_n - X$ be a well-defined random variable.

Even more important than $k = 1$ is the special case $k = 2$:

**Definition 2.14** We say that $X_n$ converges in quadratic mean to $X$, written $X_n \xrightarrow{\text{qm}} X$, if

$$\mathrm{E}\, |X_n - X|^2 \to 0 \text{ as } n \to \infty.$$

Convergence in quadratic mean is important for two reasons: First, it is often quite easy to check: In Exercise 2.1, you are asked to prove that $X_n \xrightarrow{\text{qm}} c$ if and only if $\mathrm{E}\, X_n \to c$ and $\mathrm{Var}\, X_n \to 0$ for some constant $c$. Second, quadratic mean convergence is stronger than convergence in probability, which means that weak consistency of an estimator may be established by checking that it converges in quadratic mean. This latter property is a corollary of the following result:

**Theorem 2.15** Let $k > 0$ be fixed. Then $X_n \xrightarrow{k} X$ implies $X_n \xrightarrow{P} X$.

**Proof:** The proof relies on Markov's inequality (1.22), which states that

$$P(|X_n - X| \geq \epsilon) \leq \frac{1}{\epsilon^k} \mathrm{E}\, |X_n - X|^k \tag{2.8}$$

for an arbitrary $\epsilon > 0$. If $X_n \xrightarrow{k} X$, then by definition the right hand side of inequality (2.8) goes to zero as $n \to \infty$. But this implies that $P(|X_n - X| \geq \epsilon)$ also goes to zero as $n \to \infty$. Therefore, $X_n \xrightarrow{P} X$ by definition. ■

**Example 2.16** Any unbiased estimator is consistent if its variance goes to zero. This fact follows directly from Exercise 2.1 and Theorem 2.15: If the estimator $\delta_n$ is unbiased for the parameter $\theta$, then by definition $\mathrm{E}\, \delta_n = \theta$, which is stronger than the statement $\mathrm{E}\, \delta_n \to \theta$. If we also have $\mathrm{Var}\, \delta_n \to 0$, then by Exercise 2.1 we know $\delta_n \xrightarrow{\text{qm}} \theta$. This implies $\delta_n \xrightarrow{P} \theta$.

## Exercises for Section 2.1

**Exercise 2.1** Prove that for a constant $c$, $X_n \xrightarrow{\text{qm}} c$ if and only if $\mathrm{E}\, X_n \to c$ and $\mathrm{Var}\, X_n \to 0$.

**Exercise 2.2** The converse of Theorem 2.15 is not true. Construct an example in which $X_n \xrightarrow{P} 0$ but $\mathrm{E}\, X_n = 1$ for all $n$ (by Exercise 2.1, if $\mathrm{E}\, X_n = 1$, then $X_n$ cannot converge in quadratic mean to 0).

**Hint:** The mean of a random variable may be strongly influenced by a large value that occurs with small probability (and if this probability goes to zero, then the mean can be influenced in this way without destroying convergence in probability).

**Exercise 2.3** Prove or disprove this statement: If there exists $M$ such that $P(|X_n| < M) = 1$ for all $n$, then $X_n \xrightarrow{P} c$ implies $X_n \xrightarrow{\text{qm}} c$.

**Exercise 2.4** Prove that if $0 < k < \ell$ and $\mathrm{E}\,|X|^\ell < \infty$, then $\mathrm{E}\,|X|^k < 1 + \mathrm{E}\,|X|^\ell$.

**Hint:** Use the fact that $|X|^k \le 1 + |X|^k I\{|X| > 1\}$.

**Exercise 2.5** Prove that if $\alpha > 1$, then

$$(\mathrm{E}\,|X|)^\alpha \le \mathrm{E}\,|X|^\alpha.$$

**Hint:** Use Hölder's inequality (1.26) with $p = 1/\alpha$.

**Exercise 2.6** **(a)** Prove that if $0 < k < \ell$, then $X_n \xrightarrow{\ell} X$ implies $X_n \xrightarrow{k} X$.

**Hint:** Use Exercise 2.5.

**(b)** Prove by example that the conclusion of part (a) is not true in general if $0 < \ell < k$.

**Exercise 2.7** This exercise deals with bounded in probability sequences.

**Definition 2.17** We say that $X_n$ is *bounded in probability* if $X_n = O_P(1)$, i.e., if for every $\epsilon > 0$, there exist $M$ and $N$ such that $P(|X_n| < M) > 1 - \epsilon$ for $n > N$.

**(a)** Prove that if $X_n \xrightarrow{d} X$ for some random variable $X$, then $X_n$ is bounded in probability.

**Hint:** You may use the fact that any interval of real numbers must contain a point of continuity of $F(x)$. Also, recall that $F(x) \to 1$ as $x \to \infty$.

**(b)** Prove that if $X_n$ is bounded in probability and $Y_n \xrightarrow{P} 0$, then $X_n Y_n \xrightarrow{P} 0$.

## 2.2 Consistent estimates of the mean

For a sequence of random vectors $X_1, X_2, \ldots$, we denote the $n$th sample mean by

$$\overline{X}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n X_i.$$

We begin with a formal statement of the weak law of large numbers for an independent and identically distributed sequence. Later in this section, we discuss some case in which the sequence of random vectors is not independent and identically distributed.

### 2.2.1 The weak law of large numbers

**Theorem 2.18** *Weak Law of Large Numbers (univariate version):* Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and have finite mean $\mu$. Then $\overline{X}_n \xrightarrow{P} \mu$.

The proof of Theorem 2.18 in its full generality is somewhat beyond the scope of this chapter, though it may be proved using the tools in Section 4.1. However, by tightening the assumptions a bit, a proof can be made simple. For example, if the $X_n$ are assumed to have finite variance (not a terribly restrictive assumption), the weak law may be proved in a single line: Chebyshev's inequality (1.23) implies that

$$P\left(|\overline{X}_n - \mu| \geq \epsilon\right) \leq \frac{\text{Var } \overline{X}_n}{\epsilon^2} = \frac{\text{Var } X_1}{n\epsilon^2} \to 0,$$

so $\overline{X}_n \xrightarrow{P} \mu$ follows by definition. (This is a special case of Theorem 2.15.)

**Example 2.19** If $X \sim \text{binomial}(n, p)$, then $X/n \xrightarrow{P} p$. Although we could prove this fact directly using the definition of convergence in probability, it follows immediately from the Weak Law of Large Numbers due to the fact that $X$ may be expressed as the sum of $n$ independent and identically distributed Bernoulli random variables, each with mean $p$.

### 2.2.2 Independent but not identically distributed variables

Let us know generalize the conditions of the previous section: Suppose that $X_1, X_2, \ldots$ are independent but not necessarily identically distributed, such that $\text{E } X_i = \mu$ and $\text{Var } X_i = \sigma_i^2$. When is $\overline{X}_n$ consistent for $\mu$?

Let us explore sufficient conditions for a slightly stronger result, namely $\overline{X}_n \xrightarrow{\text{qm}} \mu$. Since $\text{E } \overline{X}_n = \mu$ for all $n$, Exercise 2.1 implies that $\overline{X}_n \xrightarrow{\text{qm}} \mu$ if and only if $\text{Var } \overline{X}_n \to 0$. (note, however, that the converse is not true; see Exercise 2.8). Since

$$\text{Var } \overline{X}_n = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2, \tag{2.9}$$

we conclude that $\overline{X}_n \xrightarrow{P} \mu$ if $\sum_{i=1}^n \sigma_i^2 = o(n^2)$.

On the other hand, if instead of $\overline{X}_n$ we consider the BLUE (best linear unbiased estimator)

$$\delta_n = \frac{\sum_{i=1}^n X_i/\sigma_i^2}{\sum_{j=1}^n 1/\sigma_j^2}, \tag{2.10}$$

then as before E $\delta_n = \mu$, but

$$\text{Var } \delta_n = \frac{1}{\sum_{j=1}^n 1/\sigma_j^2}. \tag{2.11}$$

Since $n \, \text{Var } \overline{X}_n$ and $n \, \text{Var } \delta_n$ are, respectively, the arithmetic and harmonic means of the $\sigma_i^2$, the statement that Var $\delta_n \le$ Var $\overline{X}_n$ (with equality only if $\sigma_1^2 = \sigma_2^2 = \cdots$) is a restatement of the harmonic-arithmetic mean inequality. It is even possible that Var $\overline{X}_n \to \infty$ while Var $\delta_n \to 0$: As an example, take $\sigma_i^2 = i \log(i)$ for $i > 1$.

**Example 2.20** Consider the case of simple linear regression,

$$Y_i = \beta_0 + \beta_1 z_i + \epsilon_i,$$

where the $z_i$ are known covariates and the $\epsilon_i$ are independent and identically distributed with mean 0 and finite variance $\sigma^2$. If we define

$$w_i = \frac{z_i - \overline{z}}{\sum_{j=1}^n (z_j - \overline{z})^2} \quad \text{and} \quad v_i = \frac{1}{n} - \overline{z} w_i,$$

then the least squares estimators of $\beta_0$ and $\beta_1$ are

$$\hat{\beta}_{0n} = \sum_{i=1}^n v_i Y_i \quad \text{and} \quad \hat{\beta}_{1n} = \sum_{i=1}^n w_i Y_i,$$

respectively. Since E $Y_i = \beta_0 + \beta_1 z_i$, we have

$$\text{E } \hat{\beta}_{0n} = \beta_0 + \beta_1 \overline{z} - \beta_0 \overline{z} \sum_{i=1}^n w_i - \beta_1 \overline{z} \sum_{i=1}^n w_i z_i$$

and

$$\text{E } \hat{\beta}_{1n} = \beta_0 \overline{w} + \beta_1 \sum_{i=1}^n w_i z_i.$$

Note that $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i z_i = 1$. Therefore, E $\hat{\beta}_{0n} = \beta_0$ and E $\hat{\beta}_{1n} = \beta_1$, which is to say that E $\hat{\beta}_{0n}$ and E $\hat{\beta}_{1n}$ are unbiased. Therefore, by Exercise 2.1 and Theorem 2.15, a sufficient condition for the consistency of E $\hat{\beta}_{0n}$ and E $\hat{\beta}_{1n}$ is that their variances tend to zero as $n \to \infty$. Since Var $Y_i = \sigma^2$, we obtain

$$\text{Var } \hat{\beta}_{0n} = \sigma^2 \sum_{i=1}^n v_i^2 \quad \text{and} \quad \text{Var } \hat{\beta}_{1n} = \sigma^2 \sum_{i=1}^n w_i^2.$$

With $\sum_{i=1}^n w_i^2 = \left\{\sum_{j=1}^n (z_i - \overline{z})^2\right\}^{-1}$, these expressions simplify to

$$\text{Var } \hat{\beta}_{0n} = \frac{\sigma^2}{n} + \frac{\sigma^2 \overline{z}^2}{\sum_{j=1}^n (z_i - \overline{z})^2} \quad \text{and} \quad \text{Var } \hat{\beta}_{1n} = \frac{\sigma^2}{\sum_{j=1}^n (z_i - \overline{z})^2}.$$

Therefore, $\hat{\beta}_{0n}$ and $\hat{\beta}_{1n}$ are consistent if

$$\frac{\overline{z}^2}{\sum_{j=1}^n (z_i - \overline{z})^2} \quad \text{and} \quad \frac{1}{\sum_{j=1}^n (z_i - \overline{z})^2},$$

respectively, tend to zero.

### 2.2.3 Identically distributed but not independent variables

Suppose that $X_1, X_2, \ldots$ satisfy E $X_i = \mu$ but they are not independent. Then E $\overline{X}_n = \mu$ and so $\overline{X}_n$ is consistent for $\mu$ if Var $\overline{X}_n \to 0$. In this case,

$$\text{Var } \overline{X}_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}\,(X_i, X_j). \tag{2.12}$$

For dependent sequences, it is common to generalize "identically distributed" to the stronger concept of stationarity:

**Definition 2.21** The sequence $X_1, X_2, \ldots$ is said to be **stationary** if the joint distribution of $(X_i, \ldots, X_{i+k})$ does not depend on $i$ for any $i > 0$ and any $k \geq 0$.

Note that stationary implies identically distributed, and for independent sequences these two concepts are the same.

To simplify Equation (2.12), note that for a stationary sequence, Cov $(X_i, X_j)$ depends only on the "gap" $j - i$. For example, stationarity implies Cov $(X_1, X_4)$ = Cov $(X_2, X_5)$ = Cov $(X_5, X_8) = \cdots$. Therefore, Equation (2.12) becomes

$$\text{Var } \overline{X}_n \; = \; \frac{\sigma^2}{n} + \frac{2}{n^2} \sum_{k=1}^{n-1} (n-k)\,\text{Cov}\,(X_1, X_{1+k}) \tag{2.13}$$

for a stationary sequence.

**Lemma 2.22** Expression (2.13) tends to 0 as $n \to \infty$ if $\sigma^2 < \infty$ and Cov $(X_1, X_{1+k}) \to 0$ as $k \to \infty$.

**Proof:** It is clear that $\sigma^2/n \to 0$ if $\sigma^2 < \infty$. Assuming that $\operatorname{Cov}(X_1, X_{1+k}) \to 0$, select $\epsilon > 0$ and note that if $N$ is chosen so that $|\operatorname{Cov}(X_1, X_{1+k})| < \epsilon/2$ for all $n > N$, we have

$$\left| \frac{2}{n^2} \sum_{k=1}^{n-1} (n-k) \operatorname{Cov}(X_1, X_{1+k}) \right| \le \frac{2}{n} \sum_{k=1}^{N} |\operatorname{Cov}(X_1, X_{1+k})| + \frac{2}{n} \sum_{k=N+1}^{n-1} |\operatorname{Cov}(X_1, X_{1+k})| .$$

Note that the second term on the right is strictly less than $\epsilon/2$, and the first term is a constant divided by $n$, which may be made smaller than $\epsilon/2$ by choosing $n$ large enough. ■

**Definition 2.23** For a fixed nonnegative integer $m$, the sequence $X_1, X_2, \ldots$ is called $m$-dependent if the random vectors $(X_1, \ldots, X_i)$ and $(X_j, X_{j+1}, \ldots)$ are independent whenever $j - i > m$.

Any independent and identically distributed sequence is 0-dependent and stationary. Also, any stationary $m$-dependent sequence trivially satisfies $\operatorname{Cov}(X_1, X_{1+k}) \to 0$ as $k \to \infty$, so by Lemma 2.22, $\overline{X}_n$ is consistent for any stationary $m$-dependent sequence with finite variance.

## Exercises for Section 2.2

**Exercise 2.8** Give an example of an independent sequence $X_1, X_2, \ldots$ with $\operatorname{E} X_i = \mu$ such that $\overline{X}_n \xrightarrow{P} \mu$ but $\operatorname{Var} \overline{X}_n$ does not converge to 0.

**Exercise 2.9** Suppose $X_1, X_2, \ldots$ are independent and identically distributed with mean $\mu$ and finite variance $\sigma^2$. Let $Y_i = \overline{X}_i = (\sum_{j=1}^{i} X_j)/i$.

**(a)** Prove that $\overline{Y}_n = (\sum_{i=1}^{n} Y_i)/n$ is a consistent estimator of $\mu$.

**(b)** Compute the relative efficiency $e_{\overline{Y}_n, \overline{X}_n}$ of $\overline{Y}_n$ to $\overline{X}_n$, defined as $\operatorname{Var}(\overline{X}_n)/\operatorname{Var}(\overline{Y}_n)$, for $n \in \{5, 10, 20, 50, 100, \infty\}$ and report the results in a table. Note that $n = \infty$ in the table is shorthand for the limit (of the efficiency) as $n \to \infty$.

**Exercise 2.10** Let $Y_1, Y_2, \ldots$ be independent and identically distributed with mean $\mu$ and variance $\sigma^2 < \infty$. Let

$$X_1 = Y_1, \quad X_2 = \frac{Y_2 + Y_3}{2}, \quad X_3 = \frac{Y_4 + Y_5 + Y_6}{3}, \quad \text{etc.}$$

Define $\delta_n$ as in Equation (2.10).

**(a)** Show that $\delta_n$ and $\overline{X}_n$ are both consistent estimators of $\mu$.

**(b)** Calculate the relative efficiency $e_{\overline{X}_n, \delta_n}$ of $\overline{X}_n$ to $\delta_n$, defined as $\operatorname{Var}(\delta_n)/\operatorname{Var}(\overline{X}_n)$, for $n = 5, 10, 20, 50, 100$, and $\infty$ and report the results in a table.

**(c)** Using Example 1.22, give a simple expression asymptotically equivalent to $e_{\overline{X}_n,\delta_n}$. Report its values in your table for comparison. How good is the approximation for small $n$?

## 2.3 Convergence of Transformed Sequences

Many statistic estimators of interest may be written as functions of simpler statistics whose convergence properties are known. Therefore, results that describe the behavior of the transformed sequences have central importance for the study of statistical large-sample theory. Here, we survey a few of these results. Some are unsurprising to anyone familiar with sequences of real numbers, like the fact that continuous functions preserve various modes of convergence. Others, such as Slutsky's theorem or the Cramér-Wold theorem, have no analogues for non-random sequences.

We begin with some univariate results, but to appreciate the richness of the theory it is necessary to consider the more general multivariate case. For this reason, part of this section is spent on extending the earlier univariate definitions of the chapter to the multivariate case.

### 2.3.1 Continuous transformations: The univariate case

Just as they do for sequences of real numbers, continuous functions preserve convergence of sequences of random variables. We state this result formally for both convergence in probability and convergence in distribution.

**Theorem 2.24** Suppose that $f(x)$ is a continuous function.

**(a)** If $X_n \xrightarrow{P} X$, then $f(X_n) \xrightarrow{P} f(X)$.

**(b)** If $X_n \xrightarrow{d} X$, then $f(X_n) \xrightarrow{d} f(X)$.

Proving Theorem 2.24 is quite difficult; each of the statements of the theorem will require an additional result for its proof. For convergence in probability, this additional result must wait until the next chapter (Theorem 3.9), when we show convergence in probability to be equivalent to another condition involving almost sure convergence. To prove Theorem 2.24(b), we introduce a condition equivalent to convergence in distribution:

**Theorem 2.25** $X_n \xrightarrow{d} X$ if and only if $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for all bounded and continuous real-valued functions $g(x)$.

Theorem 2.25 is proved in Exercises 2.11 and 2.12. Once this theorem is established, Theorem 2.24(b) follows easily:

**Proof of Theorem 2.24(b)** Let $g(x)$ be any bounded and continuous function. By Theorem 2.25, it suffices to show that $\mathrm{E}\,[g \circ f(X_n)] \to \mathrm{E}\,[g \circ f(X)]$. Since $f(x)$ is continuous, $g \circ f(x)$ is bounded and continuous. Therefore, another use of Theorem 2.25 proves that $\mathrm{E}\,[g \circ f(X_n)] \to \mathrm{E}\,[g \circ f(X)]$ as desired. ■

### 2.3.2 Multivariate Extensions

We now extend our notions of convergence to the multivariate case. Several earlier results from this chapter, such as the weak law of large numbers and the results on continuous functions, are generalized to this case.

The multivariate definitions of convergence in probability and convergence in $k$th mean are straightforward and require no additional development:

**Definition 2.26** $\mathbf{X}_n$ converges in probability to $\mathbf{X}$ (written $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$) if for any $\epsilon > 0$,

$$P\left(\|\mathbf{X}_n - \mathbf{X}\| < \epsilon\right) \to 1 \text{ as } n \to \infty.$$

**Definition 2.27** $\mathbf{X}_n$ converges in $k$th mean to $\mathbf{X}$ (written $\mathbf{X}_n \xrightarrow{k} \mathbf{X}$) if for any $\epsilon > 0$,

$$\mathrm{E}\,\|\mathbf{X}_n - \mathbf{X}\|^k \to 0 \text{ as } n \to \infty.$$

As a special case, we have convergence in quadratic mean, written $\mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{X}$, when

$$\mathrm{E}\,\left[(\mathbf{X}_n - \mathbf{X})^T(\mathbf{X}_n - \mathbf{X})\right] \to 0 \text{ as } n \to \infty.$$

**Example 2.28** *The weak law of large numbers* As in the univariate case, we define the $n$th sample mean of a random vector sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ to be

$$\overline{\mathbf{X}}_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i.$$

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed and have finite mean $\boldsymbol{\mu}$. Then $\overline{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}$. Thus, the multivariate weak law is identical to the univariate weak law (and its proof using characteristic functions is also identical to the univariate case; see Section 4.1).

It is also straightforward to extend convergence in distribution to random vectors, though in order to do this we need the multivariate analogue of Equation (2.4), the distribution function. To this end, let

$$F(\mathbf{x}) \stackrel{\text{def}}{=} P(\mathbf{X} \leq \mathbf{x}),$$

where $\mathbf{X}$ is a random vector in $\mathbb{R}^k$ and $\mathbf{X} \leq \mathbf{x}$ means that $X_i \leq x_i$ for all $1 \leq i \leq k$.

**Definition 2.29** $\mathbf{X}_n$ converges in distribution to $\mathbf{X}$ (written $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$) if for any point $\mathbf{c}$ at which $F(\mathbf{x})$ is continuous,

$$F_n(\mathbf{c}) \to F(\mathbf{c}) \text{ as } n \to \infty.$$

There is one rather subtle way in which the multivariate situation is not quite the same as the univariate situation. In the univariate case, it is very easy to characterize the points of continuity of $F(x)$: The distribution function of the random variable $X$ is continuous at $x$ if and only if $P(X = x) = 0$. However, this simple characterization no longer holds true for random vectors; a point $\mathbf{x}$ may be a point of discontinuity yet still satisfy $P(\mathbf{X} = \mathbf{x}) = 0$. The task in Exercise 2.14 is to produce an example of this phenomenon.

As a final generalization, we extend Theorem 2.24 to the multivariate case.

**Theorem 2.30** Suppose that $\mathbf{f} : S \to \mathbb{R}^\ell$ is a continuous function defined on some subset $S \subset \mathbb{R}^k$, $\mathbf{X}_n$ is a $k$-component random vector, and $P(\mathbf{X} \in S) = 1$.

**(a)** If $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, then $\mathbf{f}(\mathbf{X}_n) \xrightarrow{P} \mathbf{f}(\mathbf{X})$.

**(b)** If $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, then $\mathbf{f}(\mathbf{X}_n) \xrightarrow{d} \mathbf{f}(\mathbf{X})$.

The proof of Theorem 2.30 is basically the same as in the univariate case, although there are some subtleties related to using multivariate distribution functions in part (b). For proving part (b), one can use the multivariate version of Theorem 2.31:

**Theorem 2.31** $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\mathrm{E}\, g(\mathbf{X}_n) \to \mathrm{E}\, g(\mathbf{X})$ for all bounded and continuous real-valued functions $g : \mathbb{R}^k \to \mathbb{R}$.

Proving Theorem 2.31 involves a few subtleties related to the use of multivariate distribution functions, but the essential idea is exactly the same as in the univariate case (see Theorem 2.25 and Exercises 2.11 and 2.12).

### 2.3.3 The Cramér-Wold Theorem

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is a sequence of random $k$-vectors. By Theorem 2.30, we see immediately that

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \text{ implies } \mathbf{a}^T\mathbf{X}_n \xrightarrow{d} \mathbf{a}^T\mathbf{X} \text{ for any } \mathbf{a} \in \mathbb{R}^k. \tag{2.14}$$

It is not clear, however, whether the converse of statment (2.14) is true. Such a converse would be useful because it would give a means for proving multivariate convergence in distribution using only univariate methods. We will see in the next subsection that multivariate convergence in distribution does *not* follow from the mere fact that each of the components converges in distribution (if you cannot stand the suspense, see Example 2.33). Yet the converse of statement (2.14) is much stronger than the statement that each component converges in distribution. Could it be true that requiring *all* linear combinations to converge in distribution is strong enough to guarantee multivariate convergence? The answer is yes:

**Theorem 2.32** *Cramér-Wold Theorem:* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\mathbf{a}^T\mathbf{X}_n \xrightarrow{d} \mathbf{a}^T\mathbf{X}$ for all $\mathbf{a} \in \mathbb{R}^k$.

Using the machinery of characteristic functions, to be presented in Section 4.1, the proof of the Cramér-Wold Theorem is immediate; see Exercise 4.3.

### 2.3.4 Slutsky's Theorem

Related to Theorem 2.24 is the question of when we may "stack" random variables to make random vectors while preserving convergence. It is here that we encounter perhaps the biggest surprise of this section: Convergence in distribution is not preserved by "stacking".

To understand what we mean by "stacking," consider that by the definition of convergence in probability,

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \text{ implies that } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} X \\ Y \end{pmatrix}. \tag{2.15}$$

Thus, two convergent-in-probability sequences $X_n$ and $Y_n$ may be "stacked" to make a vector, and this vector must still converge in probability to the vector of stacked limits. Note that the converse of (2.15) is true by Theorem 2.24 because the function $f(x, y) = x$ is a continuous function from $\mathbb{R}^2$ to $\mathbb{R}$. By induction, we can therefore stack or unstack arbitrarily many random variables or vectors without distrurbing convergence in probability. Combining this fact with Theorem 2.24 yields a useful result; see Exercise 2.18.

However, statement 2.15 does not remain true in general if $\xrightarrow{P}$ is replaced by $\xrightarrow{d}$ throughout the statement. Consider the following example.

**Example 2.33** Take $X_n$ and $Y_n$ to be independent standard normal random variables for all $n$. These distributions do not depend on $n$ at all, and we may write $X_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{d} Z$, where $Z \sim N(0, 1)$. But it is certainly not true that

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ Z \end{pmatrix}, \tag{2.16}$$

since the distribution on the left is bivariate normal with correlation 0, while the distribution on the right is bivariate normal with correlation 1.

Expression (2.16) is untrue precisely because the marginal distributions do not uniquely determine the joint distribution. However, there are certain special cases in which the marginals do determine the joint distribution. For instance, if random variables are independent, then their marginal distributions uniquely determine their joint distribution. Indeed, we can say that if $X_n \to X$ and $Y_n \to Y$, where $X_n$ is independent of $Y_n$ and $X$ is independent of $Y$, then statement (2.15) remains true when $\stackrel{P}{\to}$ is replaced by $\stackrel{d}{\to}$ (see Exercise 2.19). As a special case, the constant $c$, when viewed as a random variable, is automatically independent of any other random variable. Since $Y_n \stackrel{d}{\to} c$ is equivalent to $Y_n \stackrel{P}{\to} c$ by Theorem 2.12, it must be true that

$$X_n \stackrel{d}{\to} X \text{ and } Y_n \stackrel{P}{\to} c \text{ implies that } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \stackrel{d}{\to} \begin{pmatrix} X \\ c \end{pmatrix} \tag{2.17}$$

if $X_n$ is independent of $Y_n$ for every $n$. The content of a powerful theorem called Slutsky's Theorem is that statement (2.17) remains true even if the $X_n$ and $Y_n$ are not independent. Although the preceding discussion involves stacking only random (univariate) variables, we present Slutsky's theorem in a more general version involving stacking random vectors.

**Theorem 2.34** *Slutsky's Theorem:* For any random vectors $\mathbf{X}_n$ and $\mathbf{Y}_n$ and any constant $\mathbf{c}$, if $\mathbf{X}_n \stackrel{d}{\to} \mathbf{X}$ and $\mathbf{Y}_n \stackrel{P}{\to} \mathbf{c}$, then

$$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \stackrel{d}{\to} \begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix}.$$

A proof of Theorem 2.34 is outlined in Exercise 2.20.

Putting several of the preceding results together yields the following corollary.

**Corollary 2.35** If $\mathbf{X}$ is a $k$-vector such that $\mathbf{X}_n \stackrel{d}{\to} \mathbf{X}$, and $Y_{nj} \stackrel{P}{\to} c_j$ for $1 \le j \le m$, then

$$\mathbf{f}\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \stackrel{d}{\to} \mathbf{f}\begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix}$$

for any continuous function $\mathbf{f} : S \subset \mathbb{R}^{k+m} \to \mathbb{R}^\ell$.

It is very common practice in statistics to use Corollary 2.35 to obtain a result, then state that the result follows "by Slutsky's Theorem". In fact, there is not a unanimously held idea in the literature about what precisely "Slutsky's Theorem" refers to; some consider the Corollary itself, or particular cases of the Corollary, to be Slutsky's Theorem. These minor differences are unimportant; the common thread in any application of "Slutsky's Theorem," however it is defined, is some combination of one sequence that converges in distribution with one or more sequences that converge in probability to constants.

**Example 2.36** *Normality of the t-statistic* Let $X_1, \ldots, X_n$ be independent and identically distributed with mean $\mu$ and finite positive variance $\sigma^2$. By Example 2.10, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Suppose that $\hat{\sigma}_n^2$ is any consistent estimator of $\sigma^2$; that is, $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$. (For instance, we might take $\hat{\sigma}_n^2$ to be the usual unbiased sample variance estimator, whose asymptotic properties will be studied later.) If $Z$ denotes a standard normal random variable, Theorem 2.34 implies

$$\begin{pmatrix} \sqrt{n}(\overline{X}_n - \mu) \\ \hat{\sigma}_n^2 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \sigma Z \\ \sigma^2 \end{pmatrix}. \tag{2.18}$$

Therefore, since $f(a, b) = a/b$ is a continuous function,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\hat{\sigma}_n^2}} \xrightarrow{d} Z. \tag{2.19}$$

It is common practice to skip step (2.18), attributing equation (2.19) directly to "Slutsky's Theorem".

## Exercises for Section 2.3

**Exercise 2.11** Here we prove part of Theorem 2.25: If $X_n \xrightarrow{d} X$ and $g(x)$ is a bounded, continuous function on $\mathbb{R}$, then $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$. (This half of Theorem 2.25 is sometimes called the univariate Helly-Bray Theorem.)

Let $F_n(x)$ and $F(x)$ denote the distribution functions of $X_n$ and $X$, as usual. For $\epsilon > 0$, take $b < c$ to be constant real numbers such that $F(b) < \epsilon$ and $F(c) > 1-\epsilon$. First, we note that since $g(x)$ is continuous, it must be *uniformly continuous* on $[b, c]$: That is, for any $\epsilon > 0$ there exists $\delta > 0$ such that $|g(x)-g(y)| < \epsilon$ whenever $|x - y| < \delta$. This fact, along with the boundedness of $g(x)$, ensures that there exists a finite set of real numbers $b = a_0 < a_1 < \cdots < a_m = c$ such that:

- Each $a_i$ is a continuity point of $F(x)$.
- $F(a_0) < \epsilon$ and $F(a_m) > 1 - \epsilon$.
- For $1 \le i \le m$, $|g(x) - g(a_i)| < \epsilon$ for all $x \in [a_{i-1}, a_i]$.

**(a)** Define

$$h(x) = \begin{cases} g(a_i) & \text{if } a_{i-1} < x \le a_i \text{ for some } 1 \le i \le m. \\ 0 & \text{otherwise.} \end{cases}$$

Prove that there exists $N$ such that $|\operatorname{E} h(X_n) - \operatorname{E} h(X)| < \epsilon$ for all $n > N$.

**Hint:** Let $M$ be such that $|g(x)| < M$ for all $x \in \mathbb{R}$. Use the fact that for any random variable $Y$,

$$\operatorname{E} h(Y) = \sum_{i=1}^{m} g(a_i) P(a_{i-1} < Y \leq a_i).$$

Also, please note that we may not write $\operatorname{E} h(X_n) - \operatorname{E} h(X)$ as $\operatorname{E}[h(X_n) - h(X)]$ because it is not necessarily the case that $X_n$ and $X$ are defined on the same sample space.

**(b)** Prove that $\operatorname{E} g(X_n) \to \operatorname{E} g(X)$.

**Hint:** Use the fact that

$$\begin{aligned} |\operatorname{E} g(X_n) - \operatorname{E} g(X)| \leq \; & |\operatorname{E} g(X_n) - \operatorname{E} h(X_n)| + |\operatorname{E} h(X_n) - \operatorname{E} h(X)| \\ & + |\operatorname{E} h(X) - \operatorname{E} g(X)|. \end{aligned}$$

**Exercise 2.12** Prove the other half of Theorem 2.25:

**(a)** Let $a$ be any continuity point of $F(x)$. Let $\epsilon > 0$ be arbitrary. Show that there exists $\delta > 0$ such that $F(a - \delta) > F(a) - \epsilon$ and $F(a + \delta) < F(a) + \epsilon$.

**(b)** Show how to define continuous functions $g_1 : \mathbb{R} \to [0, 1]$ and $g_2 : \mathbb{R} \to [0, 1]$ such that for all $x \leq a$, $g_1(x) = g_2(x - \delta) = 0$ and for all $x > a$, $g_1(x + \delta) = g_2(x) = 1$. Use these functions to bound the difference between $F_n(a)$ and $F(a)$ in such a way that this difference must tend to 0.

**Exercise 2.13** To illustrate a situation that can arise in the multivariate setting that cannot arise in the univariate setting, construct an example of a sequence $(X_n, Y_n)$, a joint distribution $(X, Y)$, and a connected subset $S \in R^2$ such that

**(i)** $(X_n, Y_n) \stackrel{d}{\to} (X, Y)$;

**(ii)** every point of $R^2$ is a continuity point of the distribution function of $(X, Y)$;

**(iii)** $P[(X_n, Y_n) \in S]$ does not converge to $P[(X, Y) \in S]$.

**Hint:** Condition (ii) may be satisfied even if the distribution of $(X, Y)$ is concentrated on a line.

**Exercise 2.14** If $X$ is a univariate random variable with distribution function $F(x)$, then $F(x)$ is continuous at $c$ if and only if $P(X = c) = 0$. Prove by counterexample that this is not true if variables $X$ and $c$ are replaced by vectors $\mathbf{X}$ and $\mathbf{c}$.

**Exercise 2.15** Suppose that $X$ and $Y$ are standard normal random variables with $\mathrm{Corr}\,(X, Y) = \rho$. Construct a computer program that simulates the distribution function $F_\rho(x, y)$ of the joint distribution of $X$ and $Y$. For a given $(x, y)$, the program should generate at least 50,000 random realizations from the distribution of $(X, Y)$, then report the proportion for which $(X, Y) \le (x, y)$. (If you wish, you can also report a confidence interval for the true value.) Use your function to approximate $F_{.5}(1, 1)$, $F_{.25}(-1, -1)$, and $F_{.75}(0, 0)$. As a check of your program, you can try it on $F_0(x, y)$, whose true values are not hard to calculate directly assuming your software has the ability to evaluate the standard normal distribution function.

**Hint:** To generate a bivariate normal random vector $(X, Y)$ with covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, start with independent standard normal $U$ and $V$, then take $X = U$ and $Y = \rho U + \sqrt{1 - \rho^2} V$.

**Exercise 2.16** Adapt the method of proof in Exercise 2.11 to the multivariate case, proving half of Theorem 2.31: If $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, then $\mathrm{E}\, g(\mathbf{X}_n) \to \mathrm{E}\, g(\mathbf{X})$ for any bounded, continuous $g : \mathbb{R}^k \to \mathbb{R}$.

**Hint:** Instead of intervals $(a_{i-1}, a_i]$ as in Exercise 2.11, use small regions $\{\mathbf{x} : a_{i,j-1} < x_i \le a_{i,j} \text{ for all } i\}$ of $\mathbb{R}^k$. Make sure these regions are chosen so that their boundaries contain only continuity points of $F(\mathbf{x})$.

**Exercise 2.17** Construct a counterexample to show that Theorem 2.34 may not be strengthened by changing $Y_n \xrightarrow{P} c$ to $Y_n \xrightarrow{P} Y$.

**Exercise 2.18** **(a)** Prove that if $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is continuous and $X_{nj} \xrightarrow{P} X_j$ for all $1 \le j \le k$, then $f(\mathbf{X}_n) \xrightarrow{P} f(\mathbf{X})$.

**(b)** Taking $f(a, b) = a + b$ for simplicity, construct an example demonstrating that part (a) is not true if $\xrightarrow{P}$ is replaced by $\xrightarrow{d}$.

**Exercise 2.19** Prove that if $X_n \to X$ and $Y_n \to Y$, where $X_n$ is independent of $Y_n$ and $X$ is independent of $Y$, then

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} Y \text{ implies that } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

**Hint:** Be careful to deal with points of discontinuity: If $X_n$ and $Y_n$ are independent, what characterizes a point of discontinuity of the joint distribution?

**Exercise 2.20** Prove Slutsky's Theorem, Theorem 2.34, using the following approach:

**(a)** Prove the following lemma:

**Lemma 2.37** Let $\mathbf{V}_n$ and $\mathbf{W}_n$ be $k$-dimensional random vectors on the same sample space.

$$\text{If } \mathbf{V}_n \xrightarrow{d} \mathbf{V} \text{ and } \mathbf{W}_n \xrightarrow{P} \mathbf{0}, \text{ then } \mathbf{V}_n + \mathbf{W}_n \xrightarrow{d} \mathbf{V}.$$

**Hint:** For arbitrary $\epsilon > 0$, let $\boldsymbol{\epsilon}$ denote the $k$-vector all of whose entries are $\epsilon$. Show that for any $\mathbf{a} \in \mathbb{R}^k$,

$$\begin{aligned} P(\mathbf{V}_n \leq \mathbf{a} - \boldsymbol{\epsilon}) + P(\|\mathbf{W}_n\| < \|\boldsymbol{\epsilon}\|) - 1 \ &\leq \ P(\mathbf{V}_n + \mathbf{W}_n \leq \mathbf{a}) \\ &\leq \ P(\mathbf{V}_n \leq \mathbf{a} + \boldsymbol{\epsilon}) + P(\|\mathbf{W}_n\| \geq \|\boldsymbol{\epsilon}\|). \end{aligned}$$

If $N$ is such that $P(\|\mathbf{W}_n\| \geq \|\boldsymbol{\epsilon}\|) < \epsilon$ for all $n > N$, rewrite the above inequalities for $n > N$. If $\mathbf{a}$ is a continuity point of $F(\mathbf{v})$, what happens as $\epsilon \to 0$?

**(b)** Show how to prove Theorem 2.34 using the lemma in part (a).