# Chapter 11
# Basic Large-Sample Theory

## 11.1 Introduction

Chapters 3, 4, 5, 6, and 7 were concerned with the derivation of UMP, UMP unbiased, and UMP invariant tests. Unfortunately, the existence of such tests turned out to be restricted essentially to one-parameter families with monotone likelihood ratio, exponential families, and group families, respectively. Tests maximizing the minimum or average power over suitable classes of alternatives exist fairly generally, but are difficult to determine explicitly, and their derivation in Chapter 8 was confined primarily to situations in which invariance considerations apply.

Despite their limitations, these approaches have proved their value by application to large classes of important situations. On the other hand, they are unlikely to be applicable to complex new problems. What is needed for such cases is a simpler, less detailed, more generally applicable formulation. The development and implementation of such an approach will be the subject of the remaining chapters. It replaces optimality by asymptotic optimality obtained by embedding the actual situation in a sequence of situations of increasing sample size, and applying optimality to the limit situation. These limits tend to be of a simple type for which optimality has been established in earlier chapters.

A feature of asymptotic optimality is that it refers not to a single test but to a sequence of tests, although this distinction will often be suppressed. An important consequence is that asymptotically optimal procedures—unlike most optimal procedures in the small-sample approach—are not unique since many different sequences have the same limit. In fact, quite different methods of construction may lead to procedures which are asymptotically optimal.

The following are some specific examples to keep in mind where finite-sample considerations fail to provide optimal procedures, but for which a large-sample approach will be more successful.

**Example 11.1.1 (One-parameter families)** Suppose $X_1, \ldots, X_n$ are i.i.d. according to some family of distributions $P_\theta$ indexed by a real-valued parameter $\theta$. Then, it was mentioned after Corollary 3.4.1 that UMP tests for testing $\theta = \theta_0$ against $\theta > \theta_0$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer Texts in Statistics, https://doi.org/10.1007/978-3-030-70578-7_11

exist for all sample sizes (under weak regularity conditions) only when the distributions $P_\theta$ constitute an exponential family. For example, location models typically do not have a monotone likelihood ratio, and so UMP tests rarely exist in this situation, though the normal location model is a happy exception. On the other hand, we shall see that under weak assumptions, there generally exist tests for one-parameter families which are asymptotically UMP in a suitable sense; see Section 15.3. For example, we shall derive an asymptotically optimal one-sided test in the Cauchy location model, among others. ■

**Example 11.1.2 (Behrens–Fisher Problem)** Consider testing the equality of means for two independent samples, from normal distributions with possibly different (unknown) variances. As previously mentioned, finite-sample optimality considerations such as unbiasedness or invariance do not lead to an optimal test, even though the setting is a multiparameter exponential family. An optimal test sequence will be derived in Example 15.5.4. ■

**Example 11.1.3 (The Chi-squared Test)** Consider $n$ multinomial trials with $k+1$ possible outcomes, labeled 1 to $k+1$. Suppose $p_j$ denotes the probability of a result in the $j$th category. Let $Y_j$ denote the number of trials resulting in category $j$, so that $(Y_1, \ldots, Y_{k+1})$ has the multinomial distribution with joint density obtained in Example 2.7.2. Suppose the null hypothesis is that $p = \pi = (\pi_1, \ldots, \pi_{k+1})$. The alternative hypothesis is unrestricted and includes all $p \neq \pi$ (with $\sum_{j=1}^{k+1} p_j = 1$). The class of alternatives is too large for a UMP test to exist, nor do unbiasedness or invariance considerations rescue the problem. The usual Chi-squared test, which is based on the test statistic $Q_n$ given by

$$Q_n = \sum_{j=1}^{k+1} \frac{(Y_j - n\pi_j)^2}{n\pi_j} , \tag{11.1}$$

will be seen to possess an asymptotic maximin property; see Section 16.3. ■

**Example 11.1.4 (Nonparametric Mean)** Suppose $X_1, \ldots, X_n$ are i.i.d. from a distribution $F$ with finite mean $\mu$ and finite variance. The problem is to test $\mu = 0$. Except when $F$ is assumed to belong to a number of simple parametric families, optimal tests for the mean rarely exist. Moreover, if we assume only a second moment, it is impossible to construct reasonable tests that are of a given size (Theorem 13.4.4). But, by making a weak restriction on the family, we will see that it is possible to construct tests that are approximately level $\alpha$ and that in addition possess an asymptotic maximin property; see Sections 13.4 and 15.6. ■

In the remaining chapters, we shall consider hypothesis testing and estimation by confidence sets from a large-sample or asymptotic point of view. In this approach, exact results are replaced by approximate ones that have the advantage of both greater simplicity and generality. But, the large-sample approach is not just restricted to situations where no finite-sample optimality approach works. As the following example shows, limit theorems often provide an easy way to approximate the critical value and power of a test (whether it has any optimality properties or not).

**Example 11.1.5 (Simple versus Simple)** Suppose that $X_1, \dots, X_n$ are i.i.d. with common distribution $P$. The problem is to test the simple null hypothesis $P = P_0$ versus the simple alternative $P = P_1$. Let $p_i$ denote the density of $P_i$ with respect to a measure $\mu$. By the Neyman–Pearson Lemma, the optimal test rejects for large values of $\sum_{i=1}^{n} \log[p_1(X_i)/p_0(X_i)]$. The exact null distribution of this test statistic may be difficult to obtain since, in general, an $n$-fold integration is required. On the other hand, since the statistic takes the simple form of a sum of i.i.d. variables, large-sample approximations to the critical value and power are easily obtained from the Central Limit Theorem (Theorem 11.2.4). ■

Another application of the large-sample approach (discussed in Section 13.2) is the study of the robustness of tests when the assumptions under which they are derived do not hold. Here, asymptotic considerations have been found to be indispensable. The problem is just too complicated for the more detailed small-sample methods to provide an adequate picture. In general, two distinct types of robustness considerations arise, which may be termed robustness of validity and robustness of efficiency; this distinction has been pointed out by Tukey and McLaughlin (1963), Box and Tiao (1964), and Mosteller and Tukey (1977). For robustness of validity, the issue is whether a level $\alpha$ test retains its level and power if the parameter space is enlarged to include a wider class of distributions. For example, in testing whether the mean of a normal population is zero, we may wish to consider the validity of a test without assuming normality. However, even when a test possesses a robustness of validity, are its optimality properties preserved when the parameter space is enlarged? This question is one of robustness of efficiency (or inference robustness). In the context of the one-sample normal location model, for example, one would study the behavior of procedures (such as a one-sample $t$-test) when the underlying distribution has thicker tails than the normal, or perhaps when the observations are not assumed independent. Large-sample theory offers valuable insights into these issues, as will be seen in Section 13.2.

When neither finite-sample nor large-sample optimal procedures exist for a given problem, it becomes important to determine procedures which have at least reasonable performance characteristics. Large-sample considerations often lead to suitable definitions and methods of construction. An example of this nature that will be treated later is the problem of testing whether an i.i.d. sample is uniformly distributed or, more generally, of goodness of fit.

As the starting point of a large-sample theory of inference, we now define asymptotic analogs of the concepts of size, level of significance, confidence coefficient, and confidence level. Suppose that data $X^{(n)}$ comes from a model indexed by a parameter $\theta \in \Omega$. Typically, $X^{(n)}$ refers to an i.i.d. sample of $n$ observations, and an asymptotic approach assumes that $n \to \infty$. Of course, two-sample problems can be considered in this setup, as well as more complex data structures. Nothing is assumed about the family $\Omega$, so that the problem may be parametric or nonparametric. First, consider testing a null hypothesis $H$ that $\theta \in \Omega_H$ versus the alternative hypothesis $K$ that $\theta \in \Omega_K$, where $\Omega_H$ and $\Omega_K$ are two mutually exclusive subsets of $\Omega$. We will be studying sequences of tests $\phi_n(X^{(n)})$.

**Definition 11.1.1** For a given level $\alpha$, a sequence of tests $\{\phi_n\}$ is *pointwise asymptotically level* $\alpha$ if, for any $\theta \in \Omega_H$,

$$\limsup_{n \to \infty} E_\theta[\phi_n(X^{(n)})] \le \alpha \ . \tag{11.2}$$

Condition (11.2) guarantees that for any $\theta \in \Omega_H$ and any $\epsilon > 0$, the level of the test will be less than or equal to $\alpha + \epsilon$ when $n$ is sufficiently large. However, the condition does not guarantee the existence of an $n_0$ (independent of $\theta$) such that

$$E_\theta[\phi_n(X^{(n)})] \le \alpha + \epsilon$$

for all $\theta \in \Omega_H$ and all $n \ge n_0$. We can therefore not guarantee the behavior of the size

$$\sup_{\theta \in \Omega_H} E_\theta[\phi_n(X^{(n)})]$$

of the test, no matter how large $n$ is.

**Example 11.1.6 (Uniform versus Pointwise Convergence)** To illustrate the above point, consider the function

$$f_n(\theta) = \alpha + (1 - \alpha) \exp(-n/\theta) \ ,$$

defined for positive integers $n$ and $\theta > 0$. Then, for any $\theta > 0$, $f_n(\theta) \to \alpha$ as $n \to \infty$; that is, $f_n(\theta)$ converges to $\alpha$ pointwise in $\theta$. However, this convergence is not uniform in $\theta$ because

$$\sup_{\theta > 0} f_n(\theta) = \alpha + (1 - \alpha) \sup_{\theta > 0} \exp(-n/\theta) = 1 \ .$$

To cast this example in the context of hypothesis testing, assume $X_1, \ldots, X_n$ are i.i.d. with the exponential distribution function

$$F_\theta(t) = P_\theta\{X_i \le t\} = 1 - \exp(-t/\theta) \ .$$

Define

$$\phi_n(X_1, \ldots, X_n) = \alpha + (1 - \alpha) I\{\min(X_1, \ldots, X_n) > 1\} \ .$$

Here and throughout, the notation $I\{E\}$ denotes an *indicator* random variable that is 1 if the event $E$ occurs and is 0 otherwise. Then, $E_\theta[\phi_n(X_1, \ldots, X_n)] = f_n(\theta)$. Hence, if $\Omega_H$ is the positive real line, the test sequence $\phi_n$ satisfies (11.2), but its size is 1 for every $n$. ■

In order to guarantee the behavior of the limiting size of a test sequence, we require the following stronger condition.

**Definition 11.1.2** The sequence $\{\phi_n\}$ is *uniformly asymptotically level* $\alpha$ if

$$\limsup_{n \to \infty} \sup_{\theta \in \Omega_H} E_\theta[\phi_n(X^{(n)})] \le \alpha \ . \tag{11.3}$$

If instead of (11.3), the sequence $\{\phi_n\}$ satisfies

$$\lim_{n \to \infty} \sup_{\theta \in \Omega_H} E_\theta[\phi_n(X^{(n)})] = \alpha \ , \tag{11.4}$$

then this value of $\alpha$ is called the limiting size of $\{\phi_n\}$.

Of course, we will also study the behavior of tests under the alternative hypothesis. The following is a weak condition that we expect reasonable tests to satisfy.

**Definition 11.1.3** The sequence $\{\phi_n\}$ is *pointwise consistent in power* if, for any $\theta$ in $\Omega_K$,

$$E_\theta[\phi_n(X^{(n)})] \to 1 \tag{11.5}$$

as $n \to \infty$.

**Example 11.1.7 (One-parameter families, Example 11.1.1, continued)** Let $T_n = T_n(X_1, \ldots, X_n)$ be a sequence of statistics, with distributions depending on a real-valued parameter $\theta$. For testing $H : \theta = \theta_0$ against $K : \theta > \theta_0$, consider the tests $\phi_n$ that reject $H$ when $T_n \ge C_n$. In many applications, it will turn out that, when $\theta = \theta_0$, $n^{1/2}(T_n - \theta_0)$ has a limiting normal distribution with mean 0 and variance $\tau^2(\theta_0)$ in the sense that, for any real number $t$,

$$P_{\theta_0}\{n^{1/2}(T_n - \theta_0) \le t\} \to \Phi(t/\tau(\theta_0)) \ , \tag{11.6}$$

where $\Phi(\cdot)$ is the standard normal c.d.f. Let $z_\alpha$ satisfy $\Phi(z_\alpha) = \alpha$. Then, the test with

$$C_n = \theta_0 + \frac{\tau(\theta_0)}{n^{1/2}} z_{1-\alpha}$$

has limiting size $\alpha$, since

$$P_{\theta_0}\{T_n \ge \theta_0 + \frac{\tau(\theta_0)}{n^{1/2}} z_{1-\alpha}\} \to \alpha \ .$$

Consider next the power of $\phi_n$ under the assumption that not only (11.6) holds, but that it remains valid when $\theta_0$ is replaced by any $\theta > \theta_0$. Then, the power of $\phi_n$ against $\theta$ is

$$\beta_n(\theta) = P_\theta\{n^{1/2}(T_n - \theta) \ge z_{1-\alpha}\tau(\theta_0) - n^{1/2}(\theta - \theta_0)\}$$

and hence $\beta_n(\theta) \to 1$ for any $\theta > \theta_0$, so that the test sequence is pointwise consistent in power. ■

Similar definitions apply to the construction of confidence sets. Let $g = g(\theta)$ be the parameter function of interest, for some mapping $g$ from $\Omega$ to some space $\Omega_g$. Let $S_n = S_n(X^{(n)}) \in \Omega_g$ denote a sequence of confidence sets for $g(\theta)$.

**Definition 11.1.4** A sequence of confidence sets $S_n$ is *pointwise asymptotically level* $1 - \alpha$ if, for any $\theta \in \Omega$,

$$\liminf_{n \to \infty} P_\theta\{g(\theta) \in S_n(X^{(n)})\} \geq 1 - \alpha \ . \tag{11.7}$$

The sequence $\{S_n\}$ is uniformly asymptotically level $1 - \alpha$ if

$$\liminf_{n \to \infty} \inf_{\theta \in \Omega} P_\theta\{g(\theta) \in S_n(X^{(n)})\} \geq 1 - \alpha \ . \tag{11.8}$$

If the lim inf in the left-hand side of (11.8) can be replaced by a lim, then the left-hand side is called the limiting confidence coefficient for $\{S_n\}$.

Most of the asymptotic theory we shall consider is local in a sense that we now briefly describe. In the hypothesis testing context, any reasonable test sequence $\phi_n$ is pointwise consistent in power. However, any actual situation has finite sample size $n$ and its power against any fixed alternative is typically less than one. In order to obtain a meaningful assessment of power, one therefore considers sequences of alternatives $\theta_n$ tending to $\Omega_H$ at a suitable rate, so that the limiting power of $\phi_n$ against $\theta_n$ is less than one. (See Example 11.3.2 for a simple example of such a local approach.)

An alternative to the local approach is to consider the rate at which the power tends to one against a fixed alternative. Although there exists a large literature on this approach based on large-deviation theory, the resulting approximations tend to be less accurate and we shall not treat this topic here.

It is also important to mention that asymptotic results may provide poor approximations to the actual finite-sample setting. Furthermore, convergence to a limit as $n \to \infty$ certainly does not guarantee that the approximation will improve with increasing $n$; an example is provided by Hodges (1957). Any asymptotic result should therefore be accompanied by an investigation of its reliability for finite sample sizes. Such checks can be carried out by simulations studies or higher-order asymptotic analysis.

The concepts and definitions presented in this introduction will be explored more fully in the remaining chapters. First, we need techniques to be able to approximate significance levels, power functions, and confidence coefficients. To this end, the rest of this chapter is devoted to useful results from the theory of weak convergence and other convergence concepts.

## 11.2 Weak Convergence and Central Limit Theorems

In this section, the basic notation, definitions, and results from the theory of weak convergence are introduced. The main theorems will be presented without proof, but we will provide illustrations of their use. For a more complete background, the reader is referred to Pollard (1984), Dudley (1989), or Billingsley (1995).

Let $X$ denote a $k \times 1$ random vector (which is just a vector-valued random variable), so that the $i$th component $X_i$ of $X$ is a real-valued random variable. Then, $X^\top = (X_1, \ldots, X_k)$. The (multivariate) cumulative distribution function (c.d.f.) of $X$ is defined to be:

$$F_X(x_1, \ldots, x_k) = P\{X_1 \le x_1, \ldots, X_k \le x_k\} .$$

Here, the probability $P$ refers to the probability on whatever space $X$ is defined. A point $x^\top = (x_1, \ldots, x_k)$ at which the c.d.f. $F_X(\cdot)$ is continuous is called a *continuity point* of $F_X$. Alternatively, $x$ is a continuity point of $F_X$ if the boundary of the set of $(y_1, \ldots, y_k)$ such that $y_i \le x_i$ for all $i$ has probability 0 under the distribution of $X$.[1] As an example, the multivariate normal distribution was first studied in Section 3.9.2.

**Definition 11.2.1** A sequence of random vectors $\{X_n\}$ with c.d.f.s $\{F_{X_n}(\cdot)\}$ is said to *converge in distribution* (or *in law*) to a random vector $X$ with c.d.f. $F_X(\cdot)$ if

$$F_{X_n}(x_1, \ldots, x_k) \to F_X(x_1, \ldots, x_k)$$

at all continuity points $x^\top = (x_1, \ldots, x_k)$ of $F_X(\cdot)$. This convergence will also be denoted $X_n \overset{d}{\to} X$. Because it really only has to do with the laws of the random variables (and not with the random variables themselves), we may also equivalently say $F_{X_n}$ converges weakly to $F_X$, written $F_{X_n} \overset{d}{\to} F_X$.[2]

The limiting random vector $X$ plays an auxiliary role, since any random variable with the same distribution would serve the same purpose. Therefore, the notation will sometimes be abused so that we also say $X_n$ converges in distribution to the c.d.f. $F$, written $X_n \overset{d}{\to} F$.

[1] In general, the *boundary* of a set $E$ in $\mathbb{R}^k$, denoted by $\partial E$ is defined as follows. The closure of $E$, denoted by $\bar{E}$, is the set of $x \in \mathbb{R}^k$ for which there exists a sequence $x_n \in E$ with $x_n \to x$. The set $E$ is *closed* if $E = \bar{E}$. The *interior* of $E$, denoted by $E^\circ$, is the set of $x$ such that, for some $\epsilon > 0$, the *Euclidean ball* with center $x$ and radius $\epsilon$, defined by $\{y \in \mathbb{R}^k : |y - x| < \epsilon\}$, is contained in $E$. Here $|\cdot|$ denotes the usual Euclidean norm. The set $E$ is *open* if $E = E^\circ$. If $E^c$ denotes the complement of a set $E$, then evidently $E^\circ$ is the complement of the closure of $E^c$, and so $E$ is open if and only if $E^c$ is closed. The boundary $\partial E$ of a set $E$ is then defined to be $\bar{E} - E^\circ = \bar{E} \cap (E^\circ)^c$.

[2] The term *weak convergence* (also sometimes called weak star convergence) distinguishes this type of convergence from stronger convergence concepts to be discussed later. However, the term is used because it is a special case of convergence in the weak star topology for elements in a Banach space (such as the space of signed measures on $\mathbb{R}^k$), though we will make no direct use of any such topological notions.

There are many equivalent characterizations of weak convergence, some of which are recorded in the next theorem.

**Theorem 11.2.1 (Portmanteau Theorem)** *Suppose $X_n$ and $X$ are random vectors in $\mathbb{R}^k$. The following are equivalent:*

*(i)* $X_n \overset{d}{\to} X$.
*(ii)* $Ef(X_n) \to Ef(X)$ *for all bounded, continuous real-valued functions* $f$.
*(iii)* *For any open set* $O$ *in* $\mathbb{R}^k$, $\liminf P(X_n \in O) \geq P(X \in O)$.
*(iv)* *For any closed set* $G$ *in* $\mathbb{R}^k$, $\limsup P(X_n \in G) \leq P(X \in G)$.
*(v)* *For any set* $E$ *in* $\mathbb{R}^k$ *for which* $\partial E$*, the boundary of* $E$*, satisfies* $P(X \in \partial E) = 0$, $P(X_n \in E) \to P(X \in E)$.
*(vi)* $\liminf Ef(X_n) \geq Ef(X)$ *for any nonnegative continuous* $f$.

Another equivalent characterization of weak convergence is based on the notion of the characteristic function of a random vector.

**Definition 11.2.2** The *characteristic function* of a random vector $X$ (taking values in $\mathbb{R}^k$) is the function $\zeta_X(\cdot)$ from $\mathbb{R}^k$ to the complex plane given by

$$\zeta_X(t) = E(e^{i\langle t, X\rangle}).$$

In the definition, $\langle t, X\rangle$ refers to the usual inner product, so that

$$\langle t, X\rangle = t^T X = \sum_{j=1}^{k} t_j X_j .$$

Two important properties of characteristic functions are the following. First, the distribution of $X$ is uniquely determined by its characteristic function. Second, the characteristic function of a sum of independent real-valued random variables is the product of the individual characteristic functions (Problem 11.7).

**Example 11.2.1 (Multivariate Normal Distribution)** Suppose a random vector $X^\top = (X_1, \ldots, X_k)$ is $N(\mu, \Sigma)$, the multivariate normal distribution with mean vector $\mu^\top = (\mu_1, \ldots, \mu_k)$ and covariance matrix $\Sigma$. In the case $k = 1$, if $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$, its characteristic function is:

$$E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}\sigma} e^{[-(x-\mu)^2/2\sigma^2]} dx = \exp(it\mu - \frac{1}{2}\sigma^2 t^2) , \qquad (11.9)$$

which can be verified by a simple integration (Problem 11.8). To obtain the characteristic function for $k > 1$, note that

$$\zeta_X(t) = E(e^{i\langle t, X\rangle})$$

is the characteristic function

$$\zeta_{\langle t, X\rangle}(\lambda) = E(e^{\lambda i\langle t, X\rangle})$$

of $\langle t, X \rangle$ evaluated at $\lambda = 1$. Now if $X$ is multivariate normal $N(\mu, \Sigma)$, then $\langle t, X \rangle$ is univariate normal with mean $\langle t, \mu \rangle$ and variance $\langle \Sigma t, t \rangle = t^\top \Sigma t$. Therefore, by the case $k = 1$, we find that

$$E(e^{i\langle t, X \rangle}) = \exp(i\langle t, \mu \rangle - \frac{1}{2}\langle \Sigma t, t \rangle) \ . \blacksquare \tag{11.10}$$

**Theorem 11.2.2 (Continuity Theorem)** *$X_n \overset{d}{\to} X$ in $\mathbb{R}^k$ if and only if*

$$\zeta_{X_n}(t) \to \zeta_X(t)$$

*for all $t$ in $\mathbb{R}^k$.*

Note that it is not enough to assume $\zeta_{X_n}(t) \to \zeta(t)$ for some limit function $\zeta(\cdot)$ in order to conclude $X_n \overset{d}{\to} X$; one must know that $\zeta(\cdot)$ is the characteristic function of some random variable (or that $\zeta(\cdot)$ is continuous at 0) (Problem 11.9).

Weak convergence of random vectors on $\mathbb{R}^k$ can be reduced to studying weak convergence on the real line by means of the following result, the proof of which follows immediately from Theorem 11.2.2 (Problem 11.10).

**Theorem 11.2.3 (Cramér–Wold Device)** *A sequence of random vectors $X_n$ on $\mathbb{R}^k$ satisfies $X_n \overset{d}{\to} X$ iff $\langle t, X_n \rangle \overset{d}{\to} \langle t, X \rangle$ for every $t \in \mathbb{R}^k$.*

The following result is crucial for this and the following chapters.

**Theorem 11.2.4 (Multivariate Central Limit Theorem)** *Let $X_n^\top = (X_{n,1}, \ldots, X_{n,k})$ be a sequence of i.i.d. random vectors with mean vector $\mu^\top = (\mu_1, \ldots, \mu_k)$ and covariance matrix $\Sigma$. Let $\overline{X}_{n,j} = \frac{1}{n}\sum_{i=1}^n X_{i,j}$. Then*

$$(n^{1/2}(\overline{X}_{n,1} - \mu_1), \ldots, n^{1/2}(\overline{X}_{n,k} - \mu_k))^\top \overset{d}{\to} N(0, \Sigma) \ .$$

To cover situations in which the distribution varies with sample size, we will deal with a *triangular array* of variables $\{X_{n,i} : \ 1 \le i \le r_n, \ \ n = 1, 2, \ldots\}$, where it is assumed $r_n \to \infty$ as $n \to \infty$. Typically, $r_n = n$, and so the term triangular array is an appropriate description, but note that the term triangular array is used even if $r_n \neq n$. The following limit theorem provides sufficient conditions for asymptotic normality for a normalized sum of real-valued variables making up a triangular array. (See Billingsley (1995), p. 369.)

**Theorem 11.2.5 (Lindeberg Central Limit Theorem)** *Suppose, for each $n$, $X_{n,1}, \ldots, X_{n,r_n}$ are independent real-valued random variables. Assume $E(X_{n,i}) = 0$ and $\sigma_{n,i}^2 = E(X_{n,i}^2) < \infty$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{n,i}^2$. Suppose, for each $\epsilon > 0$,*

$$\sum_{i=1}^{r_n} \frac{1}{s_n^2} E[X_{n,i}^2 I\{|X_{n,i}| > \epsilon s_n\}] \to 0 \qquad \text{as } n \to \infty. \tag{11.11}$$

*Then, $\sum_{i=1}^{r_n} X_{n,i}/s_n \overset{d}{\to} N(0, 1)$.*

For most applications, *Lindeberg's condition* (11.11) can be verified by *Lyapounov's Condition*, which says that, for some $\delta > 0$, $|X_{n,i}|^{2+\delta}$ are integrable and

$$\lim_{n\to\infty} \sum_{i=1}^{r_n} \frac{1}{s_n^{2+\delta}} E[|X_{n,i}|^{2+\delta}] = 0 \,. \tag{11.12}$$

Indeed, (11.12) implies (11.11) (Problem 11.11), and the result may be stated as follows.

**Corollary 11.2.1 (Lyapounov Central Limit Theorem).** *Suppose, for each $n$, $X_{n,1}, \ldots, X_{n,r_n}$ are independent. Assume $E(X_{n,i}) = 0$ and $\sigma^2_{n,i} = E(X^2_{n,i}) < \infty$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma^2_{n,i}$. Suppose, for some $\delta > 0$, (11.12) holds. Then,*

$$\sum_{i=1}^{r_n} X_{n,i}/s_n \xrightarrow{d} N(0,1).$$

**Example 11.2.2 (Uniformly Bounded $2+\delta$ Moments)** Suppose, for each $n$, $X_{n,1}, \ldots, X_{n,n}$ are independent. Assume $E(X_{n,i}) = 0$ and, for some $\delta > 0$,

$$\sup_{n,i} E(|X_{n,i}|^{2+\delta}) < \infty \,.$$

Let $\sigma^2_{n,i} = E(X^2_{n,i})$ and set

$$\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma^2_{n,i} \,.$$

Assume $\bar{\sigma}_n^2 \to \sigma^2_\infty < \infty$. Then, Lyapounov's Condition (11.12) holds (Problem 11.12) and, letting $\bar{X}_n = \sum_{i=1}^n X_{n,i}/n$, we have

$$\sqrt{n}\bar{X}_n \xrightarrow{d} N(0, \sigma^2_\infty) \,.$$

The result holds even if $\sigma^2_\infty = 0$ with the interpretation that $N(0,0)$ is the distribution that is point mass at zero. ■

There also exists a partial converse to Lindeberg's Central Limit Theorem, due to Feller and Lévy. (See Billingsley (1995) , p. 574.)

**Theorem 11.2.6** *Suppose, for each $n$, $X_{n,1}, \ldots, X_{n,r_n}$ are independent, mean 0, $\sigma^2_{n,i} = E(X^2_{n,i}) < \infty$ and $s_n^2 = \sum_{i=1}^{r_n} \sigma^2_{n,i}$. Also, assume the array is uniformly asymptotically negligible; that is,*

$$\max_{1\le i\le r_n} P\{|X_{n,i}/s_n| \ge \epsilon\} \to 0 \tag{11.13}$$

*for any $\epsilon > 0$. If $\sum_{i=1}^{r_n} X_{n,i}/s_n \xrightarrow{d} N(0,1)$ , then the Lindeberg Condition (11.11) is satisfied.*

**Corollary 11.2.2** *Suppose, for each n, $X_{n,1}, \ldots, X_{n,n}$ are i.i.d. with mean 0 and variance $\sigma_n^2$. Let $s_n^2 = n\sigma_n^2$. Assume $\sum_{i=1}^n X_{n,i}/s_n \overset{d}{\to} N(0,1)$. Then, the Lindeberg Condition (11.11) is satisfied.*

Corollary 11.2.2 follows from Theorem 11.2.6 because the assumption that the $n$th row of the triangular array is i.i.d. implies the array is uniformly asymptotically negligible, so the condition (11.13) holds. Indeed,

$$P\{|X_{n,i}|/s_n \ge \epsilon\} \le \frac{E(|X_{n,i}|^2)}{s_n^2 \epsilon^2} = \frac{1}{n\epsilon^2} \to 0\,.$$

The following Berry–Esseen Theorem gives information on the error in the normal approximation provided by the Central Limit Theorem.

**Theorem 11.2.7** *Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with c.d.f. $F$. Let $\mu(F)$ denote the mean of $F$ and let $\sigma^2(F)$ denote the variance of $F$, assumed finite and nonzero. Let $S_n = \sum_{i=1}^n X_i$. Then, there exists a universal constant $C$ (not depending on $F$, $n$, or $x$) such that*

$$\left| P\left\{ \frac{S_n - n\mu(F)}{n^{1/2}\sigma(F)} \le x \right\} - \Phi(x) \right| \le \frac{C}{n^{1/2}} \frac{E_F[|X_1 - \mu(F)|^3]}{\sigma(F)^3}\,, \tag{11.14}$$

*where $\Phi(\cdot)$ denotes the standard normal c.d.f.*

The Berry–Esseen Theorem holds if $C = 0.4748$; see Shevstova (2014). The smallest value of $C$ for which the result holds is unknown, but it is known that it fails for $C < 0.4097$ (van Beek (1972)).

If $F$ is a fixed distribution with finite third moment and nonzero variance, the right side of (11.14) tends to zero and hence the left side of (11.14) tends to zero uniformly in $x$. Furthermore, if $\mathbf{F}$ is the family of distributions $F$ with

$$\frac{E_F[|X - \mu(F)|^3]}{\sigma^3(F)} < B\,, \tag{11.15}$$

for some fixed $B < \infty$, then this convergence is also uniform in $F$ as $F$ varies in $\mathbf{F}$. Thus, if $S_n$ is the sum of $n$ i.i.d. variables with distribution $F_n$ in $\mathbf{F}$, then

$$\sup_x \left| P\left\{ \frac{S_n - n\mu(F_n)}{n^{1/2}\sigma(F_n)} \le x \right\} - \Phi(x) \right| \to 0\,. \tag{11.16}$$

**Example 11.2.3** Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli trials with probability of success $p$. Then, $S_n = \sum_i X_i$ is binomial based on $n$ trials and success probability $p$, and the usual Central Limit Theorem asserts that the probability that $(S_n - np)/[np(1-p)]^{1/2}$ is less or equal to $x$ converges to $\Phi(x)$, if $p$ is not zero or one. It follows from the Berry–Esseen Theorem that this convergence is uniform in both $x$ and $p$ as long as $p \in [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$. To see why, we show that condition (11.15) is satisfied. Observe that

$$E[|X_1 - p|^3] = p(1-p)[(1-p)^2 + p^2] \le p(1-p) \ .$$

Thus,

$$E[|X_1 - p|^3]/[p(1-p)]^{3/2} \le [\epsilon(1-\epsilon)]^{-1/2} \ ,$$

so that (11.15) holds with $B^2 = \epsilon(1-\epsilon)$. Thus, (11.16) holds, so that if $S_n$ is binomial based on $n$ trials and success probability $p_n \to p \in (0,1)$, then

$$P\{\frac{S_n - np_n}{[np_n(1-p_n)]^{1/2}} \le x_n\} \to \Phi(x) \tag{11.17}$$

whenever $x_n \to x$. ■

**Example 11.2.4 (The Sample Median)** As an application of the Berry–Esseen Theorem and the previous example, the following result establishes the asymptotic normality of the sample median. Given a sample $X_1, \ldots, X_n$ with order statistics $X_{(1)} \le \cdots \le X_{(n)}$, the median $\tilde{X}_n$ is defined to be the middle-order statistic $X_{(k)}$ if $n = 2k-1$ is odd and the average of $X_{(k)}$ and $X_{(k+1)}$ if $n = 2k$ is even.

**Theorem 11.2.8** *Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with c.d.f. $F$. Assume $F(\theta) = 1/2$, and that $F$ is differentiable at $\theta$ with $F' = f$ and $f(\theta) > 0$. Let $\tilde{X}_n$ denote the sample median. Then*

$$n^{1/2}(\tilde{X}_n - \theta) \xrightarrow{d} \mathrm{N}(0, \frac{1}{4f^2(\theta)}) \ .$$

Proof. Assume first that $n$ tends to $\infty$ through odd values and, without loss of generality, that $\theta = 0$. Fix any real number $a$ and let $S_n$ be the number of $X_i$ that exceed $a/n^{1/2}$. Then the event $\{\tilde{X}_n \le a/n^{1/2}\}$ is equivalent to the event $\{S_n \le (n-1)/2\}$. But, $S_n$ is binomial with parameters $n$ and success probability $p_n = 1 - F(a/n^{1/2})$. Thus,

$$P\{n^{1/2}\tilde{X}_n \le a\} = P\{S_n \le \frac{n-1}{2}\} = P\{\frac{S_n - np_n}{[np_n(1-p_n)]^{1/2}} \le x_n\} \ ,$$

where

$$x_n = \frac{\frac{1}{2}(n-1) - np_n}{[np_n(1-p_n)]^{1/2}} = \frac{n^{1/2}(\frac{1}{2} - p_n) - 1/(2n^{1/2})}{[p_n(1-p_n)]^{1/2}} \ .$$

As $n \to \infty$, $p_n \to 1/2$ and

$$n^{1/2}(\frac{1}{2} - p_n) = a \cdot \frac{F(a/n^{1/2}) - F(0)}{a/n^{1/2}} \to af(0) \ ,$$

which implies $x_n \to 2af(0)$. Therefore, by (11.17),

$$P\{n^{1/2}\tilde{X}_n \le a\} \to \Phi[2f(0)a] ,$$

which completes the proof for odd $n$. For the case of even $n$, see Problem 11.16. ■

Another result concerning uniformity in weak convergence is the following theorem of Polyá.

**Theorem 11.2.9 (Polyá's Theorem)** *Suppose $X_n \xrightarrow{d} X$ and $X$ has a continuous c.d.f $F_X$. Let $F_{X_n}$ denote the c.d.f. of $X_n$. Then, $F_{X_n}(x)$ converges to $F_X(x)$, uniformly in $x$.*

It is interesting and important to know that weak convergence of $F_n$ to $F$ can be expressed in terms of $\rho(F_n, F)$, where $\rho$ is a metric on the space of distributions. (Some basic properties of metrics are reviewed in the appendix, Section A.2.) To be specific, on the real line, define the Lévy distance between distributions $F$ and $G$ as follows.

**Definition 11.2.3** Let $F$ and $G$ be distribution functions on the real line. The *Lévy distance* between $F$ and $G$, denoted $\rho_L(F, G)$ is defined by

$$\rho_L(F, G) = \inf\{\epsilon > 0 : \; F(x - \epsilon) - \epsilon \le G(x) \le F(x + \epsilon) + \epsilon \quad \text{for all } x\} .$$

The definition implies that $\rho_L(F, G) = \rho_L(G, F)$ and that $\rho_L$ is a metric on the space of distribution functions (Problem 11.21). Moreover, if $F_n$ and $F$ are distribution functions, then weak convergence of $F_n$ to $F$ is equivalent to $\rho_L(F_n, F) \to 0$ (Problem 11.23). In this sense, $\rho_L$ metrizes weak convergence.

We shall next consider the implication of weak convergence for the convergence of quantiles. Ideally, the $(1 - \alpha)$ quantile $x_{1-\alpha}$ of a distribution $F$ is defined by

$$F(x_{1-\alpha}) = 1 - \alpha . \tag{11.18}$$

For the solutions of (11.18), it is necessary to distinguish three cases. First, if $F$ is continuous and strictly increasing, the equation (11.18) has a unique solution. Second, if $F$ is not strictly increasing, it may happen that $F(x) = 1 - \alpha$ on an interval $[a, b)$ or $[a, b]$, so that any $x$ in such an interval could serve as a $1 - \alpha$ quantile. Then, we shall define the $1 - \alpha$ quantile as the left hand endpoint of the interval. Third, if $F$ has discontinuities, then (11.18) may have no solutions. This happens if $F(x) > 1 - \alpha$ and $\sup\{F(y) : \; y < x\} \le 1 - \alpha$, but in this case we would call $x$ the $1 - \alpha$ quantile of $F$. A general definition encompassing all these possibilities is given by

$$x_{1-\alpha} = \inf\{x : \; F(x) \ge 1 - \alpha\} . \tag{11.19}$$

This is also sometimes written as $x_{1-\alpha} = F^{-1}(1 - \alpha)$ although $F$ may not have a proper inverse function.

Weak convergence of $F_n$ to $F$ is not enough to guarantee that $F_n^{-1}(1-\alpha)$ converges to $F^{-1}(1-\alpha)$, but the following result shows this is true if $F$ is continuous and strictly increasing at $F^{-1}(1-\alpha)$.

**Lemma 11.2.1** *Let $\{F_n\}$ be a sequence of distribution functions on the real line converging weakly to a distribution function $F$. Assume $F$ is continuous and strictly increasing at $y = F^{-1}(1-\alpha)$. Then,*

$$F_n^{-1}(1-\alpha) \to F^{-1}(1-\alpha) \; .$$

PROOF. Fix $\delta > 0$. Let $y - \epsilon$ and $y + \epsilon$ be continuity points of $F$ for some $0 < \epsilon \le \delta$. Then,

$$F_n(y-\epsilon) \to F(y-\epsilon) < 1-\alpha$$

and

$$F_n(y+\epsilon) \to F(y+\epsilon) > 1-\alpha.$$

Hence, for all sufficiently large $n$,

$$y - \epsilon \le F_n^{-1}(1-\alpha) \le y + \epsilon \; ,$$

and so, $|F_n^{-1}(1-\alpha) - y| \le \delta$ for all sufficiently large $n$. Since $\delta$ was arbitrary, the result is proved. ■

The following result is of fundamental importance.

**Theorem 11.2.10 (Continuous Mapping Theorem)** *Suppose $X_n \xrightarrow{d} X$. Let $g$ be a (measurable) map from $\mathbb{R}^k$ to $\mathbb{R}^s$. Let $C$ be the set of points in $\mathbb{R}^k$ for which $g$ is continuous. If $P(X \in C) = 1$, then $g(X_n) \xrightarrow{d} g(X)$.*

**Example 11.2.5** Suppose $X_n$ is a sequence of real-valued random variables such that $X_n \xrightarrow{d} N(0, \sigma^2)$. By the Continuous Mapping Theorem, it follows that

$$\frac{X_n^2}{\sigma^2} \xrightarrow{d} \chi_1^2 \; ,$$

where $\chi_k^2$ denotes the Chi-squared distribution with $k$ degrees of freedom. More generally, suppose $X_n$ is a sequence of $k \times 1$ vector-valued random variables such that

$$X_n \xrightarrow{d} N(0, \Sigma) \; ,$$

where $\Sigma$ is assumed positive definite. Then, there exists a unique positive definite symmetric matrix $C$ such that $C \cdot C = \Sigma$ and we write $C = \Sigma^{1/2}$. (For the construction of the square root of a positive definite symmetric matrix, see Lehmann (1999), p. 306.) By the Continuous Mapping Theorem, it follows that

$$\left|C^{-1}X_n\right|^2 \xrightarrow{d} \chi_k^2 \,. \blacksquare$$

## 11.3 Convergence in Probability and Applications

As pointed out earlier, convergence in law of $X_n$ to $X$ asserts only that the distribution of $X_n$ tends to that of $X$, but says nothing about $X_n$ itself becoming close to $X$. The following stronger form of convergence provides that $X_n$ and $X$ themselves are close for large $n$.

**Definition 11.3.1** A sequence of random vectors $\{X_n\}$ *converges in probability* to $X$, written $X_n \xrightarrow{P} X$, if, for every $\epsilon > 0$,

$$P\{|X_n - X| > \epsilon\} \to 0 \qquad \text{as } n \to \infty.$$

Convergence in probability implies convergence in distribution (Problem 11.32); the converse is false in general. However, if $X_n$ converges in distribution to a distribution assigning probability one to a constant vector $c$, then $X_n$ converges in probability to $c$, and conversely. Note that, unlike weak convergence, $X_n$ and $X$ must be defined on the same probability space in order for Definition 11.3.1 to make sense.

Convergence in probability of a sequence of random vectors $X_n$ is equivalent to convergence in probability of their components. That is, if $X_n = (X_{n,1}, \ldots, X_{n,k})^\top$ and $X = (X_1, \ldots, X_k)^\top$, then $X_n \xrightarrow{P} X$ iff for each $i = 1, \ldots, k$, $X_{n,i} \xrightarrow{P} X_i$. Moreover, $X_n \xrightarrow{P} 0$ if and only if $|X_n| \xrightarrow{P} 0$ (Problem 11.33).

A sequence of real-valued random variables $X_n$ converges in probability to infinity, written $X_n \xrightarrow{P} \infty$ if, for any real number $B$,

$$P\{X_n < B\} \to 0$$

as $n \to \infty$.

The next result and the later Theorem 11.4.1 deal with the convergence of the average of i.i.d. random variables toward their expectation, and are known as the weak and strong laws of large numbers. The terminology reflects the fact that the strong law asserts a stronger conclusion than the weak law.

**Theorem 11.3.1 (Weak Law of Large Numbers)** *Let $X_i$ be i.i.d. real-valued random variables with mean $\mu$. Then,*

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu \,.$$

Note that it is possible for $\bar{X}_n$ to converge in probability to a constant even if the mean does not exist (Problem 11.29). Also, if the $X_i$ are nonnegative and the mean is not finite, then $\bar{X}_n \xrightarrow{P} \infty$ (Problem 11.36).

Suppose $X_1, \ldots, X_n$ are i.i.d. according to a model $\{P_\theta,\ \theta \in \Omega\}$. A sequence of estimators $T_n = T_n(X_1, \ldots, X_n)$ is said to be a weakly consistent (or just consistent) estimator sequence of $g(\theta)$ if, for each $\theta \in \Omega$,

$$T_n \xrightarrow{P} g(\theta) \ .$$

Thus, the consistency of an estimator sequence merely asserts convergence in probability for each value of the parameter. For example, the Weak Law of Large Numbers asserts that the sample mean is a consistent estimator of the population mean whenever the population mean exists.

**Example 11.3.1** Suppose $X_1, \ldots, X_n$ are i.i.d. according to either $P_0$ or $P_1$. If $p_i$ denotes the density of $P_i$ with respect to a dominating measure, then by the Neyman–Pearson Lemma, an optimal test rejects for large values of

$$T_n \equiv \frac{1}{n} \sum_{i=1}^{n} \log[p_1(X_i)/p_0(X_i)] \ .$$

By the Weak Law of Large Numbers, under $P_0$,

$$T_n \xrightarrow{P} -K(P_0, P_1) \ , \tag{11.20}$$

where $K(P_0, P_1)$ is the so-called Kullback–Leibler Information, defined as

$$K(P_0, P_1) = -E_{P_0}[\log(p_1(X_1)/p_0(X_1))] \ . \tag{11.21}$$

The convergence (11.20) assumes $K(P_0, P_1)$ is well-defined in the sense that the expectation in (11.21) exists. But, by Jensen's inequality (since the negative log is convex),

$$K(P_0, P_1) \geq -\log[E_{P_0}(p_1(X_1)/p_0(X_1))] \geq 0 \ .$$

If $P_0$ and $P_1$ are distinct, then the first inequality is strict, so that $K(P_0, P_1) \geq 0$ with equality iff $P_0 = P_1$. Note, however, that $K(P_0, P_1)$ may be $\infty$, but even in this case, the convergence (11.20) holds; see Problem 11.37. Similarly, under the alternative hypothesis $P_1$,

$$T_n \xrightarrow{P} E_{P_1}[\log(p_1(X_1)/p_0(X_1)] = K(P_1, P_0) \geq 0 \ .$$

Note that $K(P_0, P_1)$ need not equal $K(P_1, P_0)$.

In summary, $T_n$ converges in probability, under $P_0$, to a negative constant (possibly $-\infty$), while under $P_1$, $T_n$ converges in probability to a positive constant (assuming

$P_0$ and $P_1$ are distinct). Therefore, for testing $P_0$ versus $P_1$, the test that rejects when $T_n > 0$ is *asymptotically perfect* in the sense that both error probabilities tend to zero; that is, $P_0\{T_n > 0\} \to 0$ and $P_1\{T_n \le 0\} \to 0$. It also follows that, for fixed $\alpha \in (0, 1)$, if $\phi_n$ is a most powerful level $\alpha$ test sequence for testing $P_0$ versus $P_1$ based on $n$ i.i.d. observations, then the power of $\phi_n$ against $P_1$ tends to one. Thus, if $P_0$ and $P_1$ are fixed with $n \to \infty$, the problem is degenerate from an asymptotic point of view. ■

For convergence in probability to a constant, it is not necessary for the $X_n$ to be defined on the same probability space. Suppose $P_n$ is a probability on a probability space $(\Omega_n, \mathcal{F}_n)$, and let $X_n$ be a random vector from $\Omega_n$ to $\mathbb{R}^k$. Then, if $c$ is a fixed constant vector in $\mathbb{R}^k$, we say that $X_n$ converges to $c$ in $P_n$-probability if, for every $\epsilon > 0$,

$$P_n\{|X_n - c| > \epsilon\} \to 0 \qquad \text{as } n \to \infty\ .$$

Alternatively, we may say $X_n$ converges to $c$ in probability if it is understood that the law of $X_n$ is determined by $P_n$.

For a sequence of numbers $x_n$ and $y_n$, the notation $x_n = o(y_n)$ means $x_n/y_n \to 0$ as $n \to \infty$. For random variables $X_n$ and $Y_n$, the notation $X_n = o_P(Y_n)$ means $X_n/Y_n \overset{P}{\to} 0$. Similarly, $X_n = o_{P_n}(Y_n)$ means $X_n/Y_n \to 0$ in $P_n$-probability.

The following theorem is very useful for proving limit theorems.

**Theorem 11.3.2 (Slutsky's Theorem)** *Suppose $\{X_n\}$ is a sequence of real-valued random variables such that $X_n \overset{d}{\to} X$. Further, suppose $\{A_n\}$ and $\{B_n\}$ satisfy $A_n \overset{P}{\to} a$, and $B_n \overset{P}{\to} b$, where $a$ and $b$ are constants. Then, $A_nX_n + B_n \overset{d}{\to} aX + b$.*

PROOF. By Problem 11.34, it follows that

$$(X_n, A_n, B_n) \overset{d}{\to} (X, a, b)\ .$$

Apply the Continuous Mapping Theorem (Theorem 11.2.10). ■

The conclusion in Slutsky's Theorem may be strengthened to convergence in probability if it is assumed that $X_n \overset{P}{\to} X$. The following corollary to Slutsky's Theorem is also fundamental.

**Corollary 11.3.1** *Suppose $\{X_n\}$ is a sequence of real-valued random variables such that $X_n$ tends to $X$ in distribution, where $X$ has a cumulative distribution function $F$ which is continuous at $c$. If $C_n \to c$ in probability, then*

$$P\{X_n \le C_n\} \to F(c)\ .$$

Corollary 11.3.1 is useful even when $C_n$ are nonrandom constants tending to $c$. Also, the corollary holds even if $c = \infty$ or $c = -\infty$ (Problem 11.40), with the interpretation $F(\infty) = 1$ and $F(-\infty) = 0$.

Note that Slutsky's Theorem holds more generally if the convergence in probability assumptions are replaced by convergence in $P_n$-probability.

**Example 11.3.2 (Local Power Calculation)** Suppose $S_n$ is binomial based on $n$ trials and success probability $p$. Consider testing $p = 1/2$ versus $p > 1/2$. The uniformly most powerful test rejects for large values of $S_n$. By Example 11.2.3,

$$Z_n \equiv (S_n - \frac{n}{2})/\sqrt{n/4} \xrightarrow{d} N(0, 1) ,$$

and so the test that rejects the null hypothesis when this quantity exceeds the normal critical value $z_{1-\alpha}$ is asymptotically level $\alpha$. Let $\beta_n(p)$ denote the power of this test against a fixed alternative $p > 1/2$. Then, $(S_n - np)/\sqrt{np(1-p)}$ is asymptotically standard normal if $p$ is the true value. Hence,

$$\beta_n(p) = P_p\{Z_n > z_{1-\alpha}\} = P_p\{\frac{S_n - np}{\sqrt{np(1-p)}} > d_n(p)\} ,$$

where

$$d_n(p) = \frac{z_{1-\alpha}}{[4p(1-p)]^{1/2}} + \frac{\sqrt{n}(\frac{1}{2} - p)}{\sqrt{p(1-p)}} \to -\infty$$

if $p > 1/2$. Thus, $\beta_n(p) \to 1$ as $n \to \infty$ for any $p > 1/2$, and so the test sequence is pointwise consistent.

This result does not distinguish between alternative values of $p$. Better discrimination is obtained by considering alternatives for which the power tends to a value less than 1. This is achieved by replacing a fixed alternative $p$ by a sequence $p_n$ tending to $1/2$, so that the task of distinguishing between $1/2$ and $p_n$ becomes more difficult as information accumulates with increasing $n$. It turns out that the power will tend to a limit less than one but greater than $\alpha$ if $p_n = 1/2 + hn^{-1/2}$ if $h > 0$. To see this, note that, by Example 11.2.3, under $p_n$, $(S_n - np_n)/\sqrt{np_n(1-p_n)}$ is asymptotically standard normal. Then,

$$\beta_n(p_n) = P_{p_n}\{Z_n > z_{1-\alpha}\} = P_{p_n}\{\frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} > d_n(p_n)\} .$$

But, $d_n(p_n) \to z_{1-\alpha} - 2h$. Hence, if $Z$ denotes a standard normal variable,

$$\beta_n(p_n) \to P\{Z > z_{1-\alpha} - 2h\} = 1 - \Phi(z_{1-\alpha} - 2h) .$$

Also, note that $\beta_n(p_n) \to 1$ if $\sqrt{n}(p_n - 1/2) \to \infty$ and $\beta_n(p_n) \to \alpha$ if $\sqrt{n}(p_n - 1/2) \to 0$ (Problem 11.41). ■

The following is another useful result concerning convergence in probability.

**Theorem 11.3.3** *Suppose $X_n$ and $X$ are random vectors in $\mathbb{R}^k$ with $X_n \xrightarrow{P} X$. Let $g$ be a continuous function from $\mathbb{R}^k$ to $\mathbb{R}^s$. Then, $g(X_n) \xrightarrow{P} g(X)$.*

**Example 11.3.3 (Sample Standard Deviation)** Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with common mean $\mu$ and finite variance $\sigma^2$. The usual unbiased sample variance estimator is given by

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \,, \tag{11.22}$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean. By the Weak Law of Large Numbers, $\bar{X}_n \to \mu$ in probability and $n^{-1} \sum_{i=1}^{n} X_i^2 \to E(X_1^2) = \mu^2 + \sigma^2$ in probability. Hence,

$$\frac{n-1}{n} S_n^2 = n^{-1} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \to \sigma^2$$

in probability, by Slutsky's Theorem. Thus, $S_n^2 \to \sigma^2$ in probability, which implies $S_n \to \sigma$ in probability, by Theorem 11.3.3. ∎

**Example 11.3.4 (Confidence Intervals for a Binomial $p$)** Suppose $S_n$ is binomial based on $n$ trials and unknown success probability $p$. Let $\hat{p}_n = S_n/n$. By Example 11.2.3, for any $p \in (0, 1)$, $\sqrt{n}(\hat{p}_n - p)$ converges in distribution to $N(0, p(1-p))$. This implies $\hat{p}_n \xrightarrow{P} p$ and so

$$\sqrt{\hat{p}_n(1-\hat{p}_n)} \xrightarrow{P} \sqrt{p(1-p)}$$

as well. Therefore, by Slutsky's Theorem, for any $p \in (0, 1)$,

$$\frac{n^{1/2}(\hat{p}_n - p)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{d} N(0, 1) \,.$$

This implies that the confidence interval

$$\hat{p}_n \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \tag{11.23}$$

is pointwise consistent in level, for any fixed $p$ in $(0, 1)$, where $z_\beta$ is the $\beta$ quantile of $N(0, 1)$. Note, however, that this confidence interval is not uniformly consistent in level; in fact, for any $n$, the coverage probability can be arbitrarily close to 0 (Problem 11.42).

Unfortunately, an accumulating literature has shown that the coverage of the interval in (11.23) is quite unreliable even for large values of $n$ or $np(1-p)$, and varies quite erratically as the sample size increases. To cite just one example, the probability of the interval (11.23) covering the true $p$ when $p = 0.2$ and $1 - \alpha = 0.95$ is 0.946 when $n = 30$, and it is 0.928 when $n = 98$. This example is taken from Table 1 of Brown Cai and DasGupta (2001), who survey the literature and recommend more reliable alternatives. Because of the great practical importance of the problem, we summarize some of their principal recommendations.

For small $n$, the authors recommend two procedures. The first, which goes back to Wilson (1927) , is based on the quadratic inequality

$$|\hat{p}_n - p| \le z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \ , \tag{11.24}$$

which has probability under $p$ tending to $1-\alpha$. So, if we were testing the simple null hypothesis that $p$ is true, we can invert the test with acceptance region (11.24). Solving for $p$ in (11.24), one obtains the Wilson interval (Problem 11.43)

$$\tilde{p}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{n}}{\tilde{n}} \sqrt{\hat{p}_n \hat{q}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{4n}} \ , \tag{11.25}$$

where $\tilde{p}_n = \tilde{S}_n/\tilde{n}$, $\tilde{S}_n = S_n + \frac{1}{2} z^2_{1-\frac{\alpha}{2}}$, $\tilde{n} = n + z^2_{1-\frac{\alpha}{2}}$, and $\hat{q}_n = 1 - \hat{p}_n$. As an alternative, the authors recommend an equal-tailed Bayes interval based on the Beta prior with $a = b = 1/2$; see Example 5.7.2.

Theoretical and additional numerical support are provided in Brown, Cai and DasGupta (2002). Other approximations are reviewed in Johnson et al. (1992). ■

An immediate consequence of Slutsky's Theorem is the following. If $\tilde{T}_n \xrightarrow{d} T$ and if $E(\tilde{T}_n - T_n)^2 \to 0$, then $T_n \xrightarrow{d} T$. A less obvious but useful result is the following, due to Hajék (Problem 11.44).

**Lemma 11.3.1** *Suppose that, as $n \to \infty$,*

$$\frac{\tilde{T}_n - E(\tilde{T}_n)}{\sqrt{Var(\tilde{T}_n)}} \xrightarrow{d} T$$

*and*

$$\frac{E[(\tilde{T}_n - T_n)^2]}{Var(\tilde{T}_n)} \to 0 \ .$$

*Then,*

$$\frac{T_n - E(T_n)}{\sqrt{Var(T_n)}} \xrightarrow{d} T \ .$$

The following method is often used to prove limit theorems, especially asymptotic normality.

**Theorem 11.3.4 (Delta Method)** *Suppose $X_1, X_2, \ldots$ and $X$ are random vectors in $\mathbb{R}^k$. Assume $\tau_n(X_n - \mu) \xrightarrow{d} X$ where $\mu$ is a constant vector and $\{\tau_n\}$ is a sequence of constants $\tau_n \to \infty$.*

*(i) Suppose $g$ is a function from $\mathbb{R}^k$ to $\mathbb{R}$ which is differentiable at $\mu$ with gradient (vector of first partial derivatives) of dimension $1 \times k$ at $\mu$ equal to $\dot{g}(\mu)$.*[3] *Then,*

$$\tau_n[g(X_n) - g(\mu)] \xrightarrow{d} \dot{g}(\mu)X \ . \tag{11.26}$$

*In particular, if $X$ is multivariate normal in $\mathbb{R}^k$ with mean vector 0 and covariance matrix $\Sigma$, then*

$$\tau_n[g(X_n) - g(\mu)] \xrightarrow{d} N(0, \dot{g}(\mu)\Sigma\dot{g}(\mu)^\top) \ . \tag{11.27}$$

*(ii) More generally, suppose $g = (g_1, \ldots, g_q)^\top$ is a mapping from $\mathbb{R}^k$ to $\mathbb{R}^q$, where $g_i$ is a function from $\mathbb{R}^k$ to $\mathbb{R}$ which is differentiable at $\mu$. Let $D$ be the $q \times k$ matrix with $(i, j)$ entry equal to $\partial g_i(y_1, \ldots, y_k)/\partial y_j$ evaluated at $\mu$. Then,*

$$\tau_n[g(X_n) - g(\mu)] = \tau_n[g_1(X_n) - g_1(\mu), \ldots, g_q(X_n) - g_q(\mu)]^\top \xrightarrow{d} DX \ .$$

*In particular, if $X$ is multivariate normal in $\mathbb{R}^k$ with mean vector 0 and covariance matrix $\Sigma$, then*

$$\tau_n[g(X_n) - g(\mu)] \xrightarrow{d} N(0, D\Sigma D^\top) \ .$$

PROOF. We prove (i) with (ii) left as an exercise (Problem 11.49). Note that $X_n - \mu = o_P(1)$. Differentiability of $g$ at $\mu$ implies

$$g(x) = g(\mu) + \dot{g}(\mu)(x - \mu) + R(x - \mu) \ ,$$

where $R(y) = o(|y|)$ as $|y| \to 0$. Now,

$$\tau_n[g(X_n) - g(\mu)] - \dot{g}(\mu)\tau_n(X_n - \mu) = \tau_n R(X_n - \mu) \ .$$

By Slutsky's Theorem, it suffices to show $\tau_n R(X_n - \mu) = o_P(1)$. But,

$$\tau_n R(X_n - \mu) = \tau_n |X_n - \mu| \cdot h(X_n - \mu) \ ,$$

where $h(y) = R(y)/|y|$ and $h(0)$ is defined to be 0, so that $h$ is continuous at 0. The weak convergence hypothesis and the Continuous Mapping Theorem imply $\tau_n|X_n - \mu|$ has a limiting distribution. So, by Slutsky's Theorem, it is enough to show $h(X_n - \mu) = o_P(1)$. But, this follows by the Continuous Mapping Theorem as well. ■

Note that (11.26) and (11.27) remain true if $\dot{g}(\mu) = 0$ with the interpretation that the limit distribution places all its mass at zero, in which case we can conclude

$$\tau_n[g(X_n) - g(\mu)] \xrightarrow{P} 0 \ .$$

[3] When $k = 1$, we may also use the notation $g'(\mu)$ for the ordinary first derivative of $g$ with respect to $\mu$, as well as $g''(\mu)$ for the second derivative.

**Example 11.3.5 (Binomial Variance)** Suppose $S_n$ is binomial based on $n$ trials and success probability $p$. Let $\hat{p}_n = S_n/n$. By the Central Limit Theorem,

$$n^{1/2}(\hat{p}_n - p) \overset{d}{\to} N(0, p(1-p)) .$$

Consider estimating $g(p) = p(1-p)$. By the Delta Method,

$$n^{1/2}[g(\hat{p}_n) - g(p)] \overset{d}{\to} N(0, (1-2p)^2 p(1-p)) .$$

If $p = 1/2$, then $\dot{g}(1/2) = 0$, so that

$$n^{1/2}[g(\hat{p}_n) - g(p)] \overset{P}{\to} 0 .$$

In order to obtain a nondegenerate limit distribution in this case, note that

$$n[g(\hat{p}_n) - \frac{1}{4}] = -[n^{1/2}(\hat{p}_n - \frac{1}{2})]^2 .$$

Therefore, by the Continuous Mapping Theorem,

$$n[g(\hat{p}_n) - \frac{1}{4}] \overset{d}{\to} -X^2 ,$$

where $X$ is $N(0, 1/4)$, or

$$n[g(\hat{p}_n) - \frac{1}{4}] \overset{d}{\to} -\frac{1}{4}\chi_1^2 ,$$

where $\chi_1^2$ is a random variable distributed as Chi-squared with one degree of freedom. ■

In the case $\dot{g}(\mu) = 0$, it is not surprising that the limit distribution is a multiple of a Chi-squared variable with one degree of freedom. Indeed, suppose $k = 1$ and $g$ is twice differentiable at $\mu$ with second derivative $g''(\mu)$, so that

$$g(x) = g(\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2 + R(x-\mu) ,$$

where $R(x-\mu) = o[(x-\mu)^2]$ as $x \to \mu$. Arguing as in the proof of Theorem 11.3.4 yields

$$\tau_n^2[g(X_n) - g(\mu)] - \tau_n^2 \frac{g''(\mu)}{2}(X_n - \mu)^2 = \tau_n^2 R(X_n - \mu) = o_P(1) \qquad (11.28)$$

(Problem 11.51). By the Continuous Mapping Theorem,

$$\tau_n(X_n - \mu) \xrightarrow{d} X$$

implies

$$\tau_n^2 \frac{g''(\mu)}{2}(X_n - \mu)^2 \xrightarrow{d} \frac{g''(\mu)}{2} X^2 .$$

By Slutsky's Theorem, $\tau_n^2[g(X_n) - g(\mu)]$ has this same limiting distribution. Of course, if $X$ is $N(\mu, \sigma^2)$, then this limiting distribution is $\frac{g''(\mu)\sigma^2}{2}\chi_1^2$.

**Example 11.3.6 (Sample Correlation)** Let $(U_i, V_i)$ be i.i.d. bivariate random vectors in the plane, with both $U_i$ and $V_i$ assumed to have finite nonzero variances. Let $\sigma_U^2 = Var(U_i)$, $\sigma_V^2 = Var(V_i)$, $\mu_U = E(U_i)$, $\mu_V = E(V_i)$ and let $\rho = Cov(U_i, V_i)/(\sigma_U \sigma_V)$ be the population correlation coefficient. The usual sample correlation coefficient is given by

$$\hat{\rho}_n = \frac{\sum_{i=1}^n (U_i - \bar{U}_n)(V_i - \bar{V}_n)/n}{S_U S_V} , \tag{11.29}$$

where $\bar{U}_n = \sum U_i/n$, $\bar{V}_n = \sum V_i/n$, $S_U^2 = \sum(U_i - \bar{U}_n)^2/n$, and $S_V^2 = \sum(V_i - \bar{V}_n)^2/n$. Then, $n^{1/2}(\hat{\rho}_n - \rho)$ is asymptotically normal. The important observation is that $\hat{\rho}_n$ is a smooth function of the vector of means $\bar{X}_n$, where $X_i$ is the vector $X_i = (U_i, V_i, U_i^2, V_i^2, U_i V_i)^\top$. In fact, $\hat{\rho}_n = g(\bar{X}_n)$, where

$$g((y_1, y_2, y_3, y_4, y_5)^\top) = \frac{y_5 - y_1 y_2}{(y_3 - y_1^2)^{1/2}(y_4 - y_2^2)^{1/2}} .$$

Note that $g$ is smooth and $\dot{g}$ is readily computed. Let $\mu = E(X_i)$ denote the mean vector. Further assume that $U_i$ and $V_i$ have finite fourth moments. Then, by the multivariate CLT,

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma) ,$$

where $\Sigma$ is the covariance matrix of $X_1$. For example, the $(1, 5)$ component of $\Sigma$ is $Cov(U_1, U_1 V_1)$. Hence, by the Delta Method,

$$n^{1/2}[g(\bar{X}_n) - g(\mu)] = n^{1/2}(\hat{\rho}_n - \rho) \xrightarrow{d} N(0, \dot{g}(\mu)\Sigma\dot{g}(\mu)^\top) . \tag{11.30}$$

As an example, suppose that $(U_i, V_i)$ is bivariate normal; in this case, (11.30) reduces to (Problem 11.52)

$$n^{1/2}(\hat{\rho}_n - \rho) \xrightarrow{d} N(0, (1 - \rho^2)^2) . \tag{11.31}$$

This implies $(1 - \hat{\rho}_n^2) \xrightarrow{P} 1 - \rho^2$. Then, by Slutsky's Theorem,

$$n^{1/2}(\hat{\rho}_n - \rho)/(1 - \hat{\rho}_n^2) \xrightarrow{d} N(0, 1) ,$$

and so the confidence interval

$$\hat{\rho}_n \pm n^{-1/2} z_{1-\frac{\alpha}{2}} (1 - \hat{\rho}_n^2)$$

is a pointwise asymptotically level $1 - \alpha$ confidence interval for $\rho$. The error in this asymptotic approximation derives from both the normal approximation to the distribution of $\hat{\rho}_n$ and the fact that one is approximating the limiting variance. To counter the second of these effects, the following variance stabilization technique can be used. By the Delta Method, if $h$ is differentiable, then

$$n^{1/2}[h(\hat{\rho}_n) - h(\rho)] \xrightarrow{d} N(0, [h'(\rho)]^2 (1 - \rho^2)^2) \ .$$

The idea is to choose $h$ so that the limiting variance does not depend on $\rho$ and is a constant; such a transformation is then called a *variance stabilizing transformation*. The solution is known as Fisher's $z$-transformation and is given by

$$h(\rho) = \frac{1}{2} \log(\frac{1+\rho}{1-\rho}) = \text{arctanh}(\rho) \ .$$

Then,

$$h(\hat{\rho}_n) \pm n^{-1/2} z_{1-\frac{\alpha}{2}}$$

is a pointwise asymptotically level $1 - \alpha$ confidence interval for $h(\rho)$. The inverse function of $h$ is the hyperbolic tangent function

$$\tanh(y) = h^{-1}(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}} \ ,$$

so that

$$[\tanh(\text{arctanh}(\hat{\rho}_n) - n^{-1/2} z_{1-\frac{\alpha}{2}}), \tanh(\text{arctanh}(\hat{\rho}_n) + n^{-1/2} z_{1-\frac{\alpha}{2}})] \qquad (11.32)$$

is also a pointwise asymptotically level $1 - \alpha$ confidence interval for $\rho$.[4] ■

Sometimes, $\{X_n\}$ may not have a limiting distribution, but the weaker property of *tightness* may hold, which only requires that no probability escapes to $\pm\infty$.

**Definition 11.3.2** A sequence of random vectors $\{X_n\}$ is *tight* (or *uniformly tight*) if $\forall \epsilon > 0$, there exists a constant $B$ such that

$$\inf_n P\{|X_n| \le B\} \ge 1 - \epsilon \ .$$

[4] For discussion of this transformation, see Mudholkar (1983), Stuart and Ord, Vol. 1 (1987) and Efron and Tibshirani (1993), p. 54. Numerical evidence supports replacing $n$ by $n - 3$ in (11.32).

A bounded sequence of numbers $\{x_n\}$ is sometimes written $x_n = O(1)$; more generally $x_n = O(y_n)$ if $x_n/y_n = O(1)$. If $\{X_n\}$ is tight, we sometimes also say $X_n$ is bounded in probability, and write $|X_n| = O_P(1)$. If $X_n$ is tight and $Y_n \xrightarrow{P} 0$ (sometimes written $Y_n = o_P(1)$), then $|X_n Y_n| \xrightarrow{P} 0$ (Problem 11.64). The notation $|X_n| = O_P(|Y_n|)$ means $|X_n|/|Y_n|$ is tight.

Tightness of a sequence of random vectors in $\mathbb{R}^k$ is equivalent to each of the component variables being tight (Problem 11.45). Note that tightness, like convergence in distribution, really refers to the sequence of laws of $X_n$, denoted $\mathcal{L}(X_n)$. Thus, we shall interchangeably refer to tightness of a sequence of random variables or the sequence of their distributions.

In a statistical context, suppose $X_1, \ldots, X_n$ are i.i.d. according to a model $\{P_\theta,\ \theta \in \Omega\}$. Recall that an estimator sequence $T_n$ is a (weakly) consistent estimator of $g(\theta)$ if, for every $\theta \in \Omega$,

$$T_n - g(\theta) \to 0$$

in probability when $P_\theta$ is true. An estimator sequence $T_n$ is said to be $\tau_n$-consistent for $g(\theta)$ if, for every $\theta \in \Omega$,

$$\tau_n[T_n - g(\theta)]$$

is tight when $P_\theta$ is true. For example, if the underlying population has a finite variance, it follows from the Central Limit Theorem that the sample mean is a $n^{1/2}$-consistent estimator of the population mean.

Whenever $X_n$ converges in distribution to a limit distribution, then $\{X_n\}$ is tight, and the following partial converse is true. Just as any bounded sequence of real numbers has a subsequence which converges, so does any sequence of random variables $X_n$ that is $O_P(1)$. This important result is stated next.

**Theorem 11.3.5 (Prohorov's Theorem)** *Suppose $\{X_n\}$ is tight on $\mathbb{R}^k$. Then, there exists a subsequence $n_j$ and a random vector $X$ such that $X_{n_j} \xrightarrow{d} X$.*

## 11.4 Almost Sure Convergence

On occasion, we shall utilize a form of convergence of $X_n$ to $X$ stronger than convergence in probability.

**Definition 11.4.1** Suppose $X_n$ and $X$ are random vectors in $\mathbb{R}^k$, defined on a common probability space $(\mathcal{X}, \mathcal{F})$. Then, $X_n$ is said to *converge almost surely* (a.s.) to $X$ if $X_n(\omega) \to X(\omega)$ on a set of points $\omega$ which has probability one; that is, if

$$P\{\omega \in \mathcal{X} : \lim_{n\to\infty} |X_n(\omega) - X(\omega)| = 0\} = 1 \ .$$

This is denoted by $X_n \to X$ a.s..

Equivalently, we say that $X_n$ converges to $X$ with probability one, since there is a set of outcomes $\omega$ having probability one such that $X_n(\omega) \to X(\omega)$. If $X_n$ converges almost surely to $X$, then $X_n$ converges in probability to $X$, but the converse is false (but see Problem 11.72). Indeed, convergence in probability does not even guarantee $X_n(\omega) \to X(\omega)$ for any outcome $\omega$. The following provides a classic counterexample.

**Example 11.4.1 (Convergence in probability, but not a.s.)** Suppose $U$ is uniformly distributed on $[0, 1)$, so that $\mathcal{X}$ is [0,1), $\mathcal{F}$ is the class of Borel sets, $U = U(\omega) = \omega$, and $P$ is the uniform probability measure. For $m = 1, 2, \ldots$ and $j = 1, \ldots, m$, let $Y_{m,j}$ be one if $U \in [(j-1)/m, j/m)$ and zero otherwise. For any $m$, exactly one of the $Y_{m,j}$ is one and the rest are zero; also, $P\{Y_{m,j} = 1\} = 1/m \to 0$ as $m \to \infty$. String together all the variables so that $X_1 = Y_1$, $X_2 = Y_{2,1}$, $X_3 = Y_{2,2}$, $X_4 = Y_{3,1}$, $X_5 = Y_{3,2}$, etc. Then, $X_n \to 0$ in probability. But $X_n$ does not converge to 0 for any outcome $U$ since $X_n$ oscillates infinitely often between 0 and 1. ■

**Theorem 11.4.1 (Strong Law of Large Numbers)** *Let $X_i$ be i.i.d. real-valued random variables with mean $\mu$. Then*

$$\bar{X}_n \equiv \tfrac{1}{n} \sum_{i=1}^{n} X_i \to \mu \quad \text{a.s.}$$

*Conversely, if $\overline{X}_n \to \mu$, a.s. with $|\mu| < \infty$, then $E|X_1| < \infty$.*

In a statistical context, suppose $X_1, \ldots, X_n$ are i.i.d. according to a model $\{P_\theta,\ \theta \in \Omega\}$. Suppose, under each $\theta$, $T_n = T_n(X_1, \ldots, X_n)$ converges almost surely to $g(\theta)$. Then, $T_n$ is said to be strongly consistent estimator of $g(\theta)$.

One of the most fundamental examples of almost sure convergence is provided by the Glivenko–Cantelli Theorem. To state the result, first define the Kolmogorov–Smirnov distance between c.d.f.s $F$ and $G$ as

$$d_K(F, G) = \sup_t |F(t) - G(t)| \ . \tag{11.33}$$

**Theorem 11.4.2 (Glivenko–Cantelli Theorem)** *Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with c.d.f. $F$. Let $\hat{F}_n$ be the empirical c.d.f. defined by*

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \le t\} \ . \tag{11.34}$$

*Then,*

$$d_K(\hat{F}_n, F) \to 0 \quad a.s.$$

To prove the Glivenko–Cantelli Theorem, note that, for every fixed $t$, $\hat{F}_n(t) \to F(t)$ almost surely, by the Strong Law of Large Numbers. That this convergence is uniform in $t$ follows from the fact that $F$ is monotone (Problem 11.62).

**Example 11.4.2 (Kolmogorov–Smirnov Test)** The Glivenko–Cantelli Theorem 11.4.2 forms the basis for the Kolmogorov–Smirnov goodness of fit test, previously introduced in Section 6.13. Specifically, consider the problem of testing the simple null hypothesis that $F = F_0$ versus $F \neq F_0$. The Glivenko-Cantelli Theorem implies that, under $F$,

$$d_K(\hat{F}_n, F_0) \to d_K(F, F_0) \quad a.s.$$

(and hence in probability as well), where the right side is zero if and only if $F = F_0$. Thus, the statistic $d_K(\hat{F}_n, F_0)$ tends to be small under the null hypothesis and large under the alternative. In order for this statistic to have a nondegenerate limit distribution under $F_0$, we normalize by multiplication of $n^{1/2}$ and the Kolmogorov–Smirnov goodness of fit test statistic is given by

$$T_n \equiv \sup_{t \in \mathbb{R}} n^{1/2} |\hat{F}_n(t) - F_0(t)| = n^{1/2} d_K(\hat{F}_n, F_0) \ . \tag{11.35}$$

The Kolmogorov–Smirnov test rejects the null hypothesis if $T_n > s_{n,1-\alpha}$, where $s_{n,1-\alpha}$ is the $1 - \alpha$ quantile of the null distribution of $T_n$ when $F_0$ is the uniform $U(0, 1)$ distribution. Recall from Section 6.13 that the finite sampling distribution of $T_n$ under $F_0$ is the same for all continuous $F_0$ (also see Problem 11.68), but its exact form is difficult to express. Some approaches to obtaining this distribution are discussed in Durbin (1973) and Section 4.3 of Gibbons and Chakraborti (1992). Values for $s_{n,1-\alpha}$ have been tabled in Birnbaum (1952). For exact power calculations in both the continuous and discrete case, see Niederhausen (1981) and Gleser (1985).

By the duality of tests and confidence regions, the Kolmogorov–Smirnov test can be inverted to yield uniform confidence bands for $F$, given by

$$R_{n,1-\alpha} = \{F : \ n^{1/2} \sup_t |\hat{F}_n(t) - F(t)| \leq s_{n,1-\alpha}\} \ . \tag{11.36}$$

By construction, $P_F\{F \in R_{n,1-\alpha}\} = 1 - \alpha$ if $F$ is continuous; furthermore, the confidence band is conservative if $F$ is not continuous (Problem 11.69).

The limiting behavior of $T_n$ will be discussed in Section 16.2. In fact, when $F = F_0$, $T_n$ has a continuous strictly increasing limiting distribution with $1 - \alpha$ quantile $s_{1-\alpha}$ (and so $s_{n,1-\alpha} \to s_{1-\alpha}$). It follows that the width of the band (11.36) is $O(n^{-1/2})$. Alternatives to the Kolmogorov–Smirnov bands that are more narrow in the tails and wider in the middle are discussed in Owen (1995). ■

The following useful inequality, which holds for finite sample sizes, actually implies the Glivenko–Cantelli Theorem (Problem 11.70).

**Theorem 11.4.3 (Dvoretzky, Kiefer, Wolfowitz Inequality)** *Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with c.d.f. $F$. Let $\hat{F}_n$ be the empirical c.d.f. (11.34). Then, for any $d > 0$ and any positive integer $n$,*

$$P\{d_K(\hat{F}_n, F) > d\} \leq C \exp(-2nd^2) \ , \tag{11.37}$$

*where $C$ is a universal constant.*

Massart (1990) shows that we can take $C = 2$, which greatly improves the original value obtained by Dvoretzy et al. (1956).

**Example 11.4.3 (Monte Carlo Simulation)** Suppose $X_1, \ldots, X_n$ are i.i.d. observations with common distribution $P$. Assume $P$ is known. The problem is to determine the distribution or quantile of some real-valued statistic $T_n(X_1, \ldots, X_n)$ for a fixed finite sample size $n$. Denote this distribution by $J_n(t)$, so that

$$J_n(t) = P\{T_n(X_1, \ldots, X_n) \le t\} \ .$$

This distribution may not have a tractable form or may not be explicitly computable, but the following simulation scheme allows the distribution $J(t)$ to be estimated to any desired level of accuracy. For $j = 1, \ldots, B$, let $X_{j,1}, \ldots, X_{j,n}$ be a sample of size $n$ from $P$; then, one simply evaluates $T_n(X_{j,1}, \ldots, X_{j,n})$, and the empirical distribution of these $B$ values serves as an approximation to the true sampling distribution $J_n(t)$. Specifically, $J_n(t)$ is approximated by

$$\hat{J}_{n,B}(t) = B^{-1} \sum_{j=1}^{B} I\{T_n(X_{j,1}, \ldots, X_{j,n}) \le t\} \ .$$

For large $B$, $\hat{J}_{n,B}(t)$ will be a good approximation to the true sampling distribution $J_n(t, P)$. One (though perhaps crude) way of quantifying the closeness of this approximation is the following. By the Dvoretsky, Kiefer, Wolfowitz inequality (11.37) (with $B$ now taking over the role of $n$), there exists a universal constant $C$ so that

$$P\{d_K(\hat{J}_{n,B}, J_n) > d\} \le C \ \exp(-2Bd^2).$$

Hence, if we desire the probability of the supremum distance between $\hat{J}_{n,B}(\cdot)$ and $J_n(\cdot, P)$ to be greater than $d$ with probability less than $\epsilon$, all we need to do is ensure that $B$ is large enough so that $C \exp(-2Bd^2) \le \epsilon$. Since $B$, the number of simulations, is determined by the statistician (assuming enough computing power), the desired accuracy can be obtained. Further results on the choice of $B$ are given in Jockel (1986).

Here, we are tacitly assuming that one can easily accomplish the sampling of observations from $P$. Of course, when $P$ corresponds to a cumulative distribution function $F$ on the real line, one can usually just obtain observations from $F$ by $F^{-1}(U)$, where $U$ is a random variable having the uniform distribution on (0, 1). This construction assumes an ability to calculate an inverse function $F^{-1}(\cdot)$. A sample $X_{j,1}, \ldots, X_{j,n}$ of $n$ i.i.d. $F$ variables can then be obtained from $n$ i.i.d. Uniform (0, 1) observations $U_{j,1}, \ldots, U_{j,n}$ by the prescription $X_{j,n} = F^{-1}(U_{j,n})$. If $F^{-1}$ is not tractable, other methods for generating observations with prescribed distributions are available in statistical software packages, such as R, Excel, or Maple.

Note, however, that we have ignored any error from the use of a pseudo-random number generator, which presumably would be needed to generate the Uniform (0, 1) variables. The above idea forms the basis of many approximation schemes; for some general references on Monte Carlo simulation, see Devroye (1986) and Ripley (1987). ■

Almost sure convergence is the strongest type of convergence we have introduced and it has many consequences. For example, suppose $X_n \to X$ almost surely and $|X_n| \leq 1$ with probability one. Then, $|X| \leq 1$ with probability one, and so $E(|X|) \leq 1$; by the Lebesgue-dominated convergence theorem (Theorem 2.2.2), it follows that $E(X_n) \to E(X)$. If the assumption that $X_n \to X$ almost surely is replaced by the weaker condition that $X_n$ converges in distribution to $X$, then the argument to show $E(X_n) \to E(X)$ breaks down. However, we shall now show that the result continues to hold since the conclusion pertains only to distributional properties of $X_n$ and $X$. The argument is based on the following theorem.

**Theorem 11.4.4 (Almost Sure Representation Theorem)** *Suppose $X_n \overset{d}{\to} X$ in $\mathbb{R}^k$. Then, there exist random vectors $\widetilde{X}_n$ and $\widetilde{X}$ defined on some common probability space such that $\widetilde{X}_n$ has the same distribution as $X_n$ and $\widetilde{X}_n \to \widetilde{X}$ a.s. (and so $\tilde{X}$ has the same distribution as $X$).*

**Example 11.4.4 (Convergence of Moments)** Suppose $X_n$ and $X$ are real-valued random variables and $X_n \overset{d}{\to} X$. If the $X_n$ are uniformly bounded, then $E(X_n) \to E(X)$. To see why, construct $\tilde{X}_n$ and $\tilde{X}$ by the Almost Sure Representation Theorem and then apply the Dominated Convergence Theorem (Theorem 2.2.2) to the $\tilde{X}_n$ to conclude

$$E(X_n) = E(\tilde{X}_n) \to E(\tilde{X}) = E(X)\ . \tag{11.38}$$

If the $X_n$ are not uniformly bounded, but $X_n \geq 0$, then by Fatou's Lemma (Theorem 2.2.1), we may conclude

$$E(X) = E(\tilde{X}) \leq \liminf_n E(\tilde{X}_n) = \liminf_n E(X_n)\ .$$

As a final result, suppose $X_n \overset{d}{\to} X$ and $|X|$ has distribution $F$ which is continuous at $t$. Then, by the Continuous Mapping Theorem,

$$|X_n| I\{|X_n| \leq t\} \overset{d}{\to} |X| I\{|X| \leq t\}\ .$$

By (11.38), we may conclude

$$E[|X_n| I\{|X_n| \leq t\}] \to E[|X| I\{|X| \leq t\}]\ . \tag{11.39}$$

If, in addition, $E|X_n| \to E|X|$, then

$$E[|X_n| I\{|X_n| > t\}] \to E[|X| I\{|X| > t\}]\ . ■ \tag{11.40}$$

More generally, convergence of moments (which are not truncated) holds under uniform integrability, which we now define.

**Definition 11.4.5** A sequence of random variables $X_1, X_2, \ldots$ is called *uniformly integrable* if, given any $\epsilon > 0$, there exists $\lambda < \infty$ such that

$$\sup_n E[|X_n| I\{|X_n| > \lambda\}] < \epsilon \ .$$

The sequence is $\{X_n\}$ is called *asymptotically uniformly integrable* if

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} E[|X_n| I\{|X_n| > \lambda\}] = 0 \ .$$

Uniform integrability is slightly stronger than asymptotic uniform integrability. A sufficient condition for uniform integrability of $\{X_n\}$ is, for some $\delta > 0$, $\sup_n E(|X_n|^{1+\delta}) < \infty$. A useful result is the following.

**Theorem 11.4.5** *Suppose $X_n \xrightarrow{d} X$ and $\{X_n\}$ is asymptotically uniformly integrable. Then, $E(X_n) \to E(X)$.*

## 11.5 Problems

### *Section 11.1*

**Problem 11.1** For each $\theta \in \Omega$, let $f_n(\theta)$ be a real-valued sequence. We say $f_n(\theta)$ converges uniformly (in $\theta$) to $f(\theta)$ if

$$\sup_{\theta \in \Omega} |f_n(\theta) - f(\theta)| \to 0$$

as $n \to \infty$. If $\Omega$ if a finite set, show that the pointwise convergence $f_n(\theta) \to f(\theta)$ for each fixed $\theta$ implies uniform convergence. However, show the converse can fail even if $\Omega$ is countable.

### *Section 11.2*

**Problem 11.2** For a univariate c.d.f. $F$, show that the set of points of discontinuity is countable.

**Problem 11.3** Let $X$ be $N(0, 1)$ and $Y = X$. Determine the set of continuity points of the bivariate distribution of $(X, Y)$.

**Problem 11.4** Show that $x = (x_1, \ldots, x_k)^\top$ is a continuity point of the distribution $F_X$ of $X$ if the boundary of the set of $(y_1, \ldots, y_k)$ such that $y_i \le x_i$ for all $i$ has probability 0 under the distribution of $X$. Show by example that it is not sufficient for $x$ to have probability 0 under $F_X$ in order for $x$ to be a continuity point.

**Problem 11.5** Prove the equivalence of (i) and (vi) in the Portmanteau Theorem (Theorem 11.2.1).

**Problem 11.6** Suppose $X_n \overset{d}{\to} X$. Show that $Ef(X_n)$ need not converge to $Ef(X)$ if $f$ is unbounded and continuous, or if $f$ is bounded but discontinuous.

**Problem 11.7** Show that the characteristic function of a sum of independent real-valued random variables is the product of the individual characteristic functions. (The converse is false; counterexamples are given in Romano and Siegel (1986), Examples 4.29–4.30.)

**Problem 11.8** Verify (11.9).

**Problem 11.9** Let $X_n$ have characteristic function $\zeta_n$. Find a counterexample to show that it is not enough to assume $\zeta_n(t)$ converges (pointwise in $t$) to a function $\zeta(t)$ in order to conclude that $X_n$ converges in distribution.

**Problem 11.10** Show that Theorem 11.2.3 follows from Theorem 11.2.2.

**Problem 11.11** Show that Lyapounov's Central Limit Theorem (Corollary 11.2.1) follows from the Lindeberg Central Limit Theorem (Theorem 11.2.5).

**Problem 11.12** In Example 11.2.2, show that Lyapounov's Condition holds.

**Problem 11.13** Suppose $X_k$ is a noncentral Chi-squared variable with $k$ degrees of freedom and noncentrality parameter $\delta_k^2$.
(i) Show that $(X_k - k)/(2k)^{1/2} \overset{d}{\to} N(\mu, 1)$ if $\delta_k^2/(2k)^{1/2} \to \mu$ as $k \to \infty$.
(ii) If $c_{k,1-\alpha}$ is the $1-\alpha$ quantile of the Chi-squared distribution with $k$ degrees of freedom, deduce that $(c_{n,1-\alpha} - k)/\sqrt{2k} \to z_{1-\alpha}$.

**Problem 11.14** Suppose $X_{n,1}, \ldots, X_{n,n}$ are i.i.d. Bernoulli trials with success probability $p_n$. If $p_n \to p \in (0, 1)$, show that

$$n^{1/2}[\bar{X}_n - p_n] \overset{d}{\to} N(0, p(1-p)) .$$

Is the result true even if $p$ is 0 or 1?

**Problem 11.15** Let $X_1, \ldots, X_n$ be i.i.d. with density $p_0$ or $p_1$, and consider testing the null hypothesis $H$ that $p_0$ is true. The MP level-$\alpha$ test rejects when $\Pi_{i=1}^n r(X_i) \ge C_n$, where $r(X_i) = p_i(X_i)/p_0(X_i)$, or equivalently when

$$\frac{1}{\sqrt{n}} \left\{ \sum \log r(X_i) - E_0[\log r(X_i)] \right\} \ge k_n. \tag{11.41}$$

(i) Show that, under $H$, the left side of (11.41) converges in distribution to $N(0, \sigma^2)$ with $\sigma^2 = \text{Var}_0[\log r(X_i)]$, provided $\sigma < \infty$.
(ii) From (i) it follows that $k_n \to \sigma z_{1-\alpha}$, where $z_\alpha$ is the $\alpha$ quantile of $N(0, 1)$.
(iii) The power of the test (11.41) against $p_1$ tends to 1 as $n \to \infty$. *Hint*: Use Problem 3.41(iv).

**Problem 11.16** Complete the proof of Theorem 11.2.8 by considering $n$ even.

**Problem 11.17** Generalize Theorem 11.2.8 to the case of the $p$th sample quantile.

**Problem 11.18** Let $X_1, \ldots, X_n$ be i.i.d. normal with mean $\theta$ and variance 1. Let $\bar{X}_n$ be the usual sample mean and let $\tilde{X}_n$ be the sample median. Let $p_n$ be the probability that $\bar{X}_n$ is closer to $\theta$ than $\tilde{X}_n$ is. Determine $\lim_{n\to\infty} p_n$.

**Problem 11.19** Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with c.d.f. $F$. Assume $\exists \theta_1 < \theta_2$ such that $F(\theta_1) = 1/4$, $F(\theta_2) = 3/4$, and $F$ is differentiable, with density $f$ taking positive values at $\theta_1$ and $\theta_2$. Show that the sample inter-quartile range (defined as the difference between the 0.75 quantile and 0.25 quantile) is a $\sqrt{n}$- consistent estimator of the population inter-quartile range $(\theta_2 - \theta_1)$.

**Problem 11.20** Prove Polyá's Theorem 11.2.9. *Hint:* First consider the case of distributions on the real line.

**Problem 11.21** Show that $\rho_L(F, G)$ defined in Definition 11.2.3 is a metric; that is, show $\rho_L(F, G) = \rho_L(G, F)$, $\rho_L(F, G) = 0$ if and only if $F = G$, and

$$\rho_L(F, G) \le \rho_L(F, H) + \rho_L(H, G) .$$

**Problem 11.22** For cumulative distribution functions $F$ and $G$ on the real line, define the Kolmogorov–Smirnov distance between $F$ and $G$ to be

$$d_K(F, G) = \sup_x |F(x) - G(x)| .$$

Show that $d_K(F, G)$ defines a metric on the space of distribution functions; that is, show $d_K(F, G) = d_K(G, F)$, $d_K(F, G) = 0$ implies $F = G$ and

$$d_K(F, G) \le d_K(F, H) + d_K(H, G) .$$

Also, show that $\rho_L(F, G) \le d_K(F, G)$, where $\rho_L$ is the Lévy metric. Construct a sequence $F_n$ such that $\rho_L(F_n, F) \to 0$ but $d_K(F_n, F)$ does not converge to zero.

**Problem 11.23** Let $F_n$ and $F$ be c.d.f.s on $\mathbb{R}$. Show that weak convergence of $F_n$ to $F$ is equivalent to $\rho_L(F_n, F) \to 0$, where $\rho_L$ is the Lévy metric.

**Problem 11.24** Suppose $F$ and $G$ are two probability distributions on $\mathbb{R}^k$. Let $\mathcal{L}$ be the set of (measurable) functions $f$ from $\mathbb{R}^k$ to $\mathbb{R}$ satisfying $|f(x) - f(y)| \le |x - y|$

and $\sup_x |f(x)| \le 1$, where $| \cdot |$ is the usual Euclidean norm. Define the Bounded-Lipschitz Metric as

$$\lambda(F, G) = \sup\{|E_F f(X) - E_G f(X)| : \ f \in \mathcal{L}\} .$$

Show that $F_n \overset{d}{\to} F$ is equivalent to $\lambda(F_n, F) \to 0$. Thus, weak convergence on $\mathbb{R}^k$ is metrizable. [See examples 21–22 in Pollard (1984).]

**Problem 11.25** For a c.d.f. $F$ with quantile function defined by

$$F^{-1}(u) = \inf\{x : \ F(x) \ge u\} ,$$

show that: (i) $F(x) \ge u$ is equivalent to $F^{-1}(u) \le x$.
(ii) $F^{-1}(\cdot)$ is nondecreasing and left continuous with right-hand limits.
(iii) $F(F^{-1}(u)) \le u$ with equality if $F$ is continuous at $F^{-1}(u)$.

**Problem 11.26** (i) Construct a sequence of distribution functions $\{F_n\}$ on the real line such that $F_n \overset{d}{\to} F$, but the convergence $F_n^{-1}(1-\alpha) \to F^{-1}(1-\alpha)$ fails, even if $F$ is assumed continuous. (ii) On the other hand, if $F$ is assumed continuous (but not necessarily strictly increasing), show that

$$F_n(F_n^{-1}(1-\alpha)) \to F(F^{-1}(1-\alpha)) = 1-\alpha .$$

[Note the left side need not be $1-\alpha$ since $F_n$ is not assumed continuous.]

## *Section 11.3*

**Problem 11.27** (Markov's Inequality) Let $X$ be a real-valued random variable with $X \ge 0$. Show that, for any $t > 0$,

$$P\{X \ge t\} \le \frac{E[XI\{X \ge t\}]}{t} \le \frac{E(X)}{t} ;$$

here $I(X \ge t)$ is the indicator variable that is 1 if $X \ge t$ and is 0 otherwise.

**Problem 11.28** (Chebyshev's Inequality) (i) Show that, for any real-valued random variable $X$ and any constants $a > 0$ and $c$,

$$E(X-c)^2 \ge a^2 P\{|X-c| \ge a\} .$$

(ii) Hence, if $X_n$ is any sequence of random variables and $c$ is a constant such that $E(X_n - c)^2 \to 0$, then $X_n \to c$ in probability. Give a counterexample to show the converse is false.

**Problem 11.29** Give an example of an i.i.d. sequence of real-valued random variables such that the sample mean converges in probability to a finite constant, yet the mean of the sequence does not exist.

**Problem 11.30** Prove the following generalization of Lemma 11.2.1. Suppose $\{\hat{F}_n\}$ is a sequence of random distribution functions satisfying $\hat{F}_n(x) \xrightarrow{P} F(x)$ at all $x$ which are continuity points of a fixed distribution function $F$. Assume $F$ is continuous and strictly increasing at $F^{-1}(1-\alpha)$. Then,

$$\hat{F}_n^{-1}(1-\alpha) \xrightarrow{P} F^{-1}(1-\alpha) \ .$$

**Problem 11.31** Prove a result analogous to Problem 11.26 if $\{\hat{F}_n\}$ is a random sequence, similar to how Problem 11.30 is a generalization of Lemma 11.2.1.

**Problem 11.32** Suppose $X_n$ and $X$ are real-valued random variables (defined on a common probability space). Prove that, if $X_n$ converges to $X$ in probability, then $X_n$ converges in distribution to $X$. Show by counterexample that the converse is false. However, show that if $X$ is a constant with probability one, then $X_n$ converging to $X$ in distribution implies $X_n$ converges to $X$ in probability.

**Problem 11.33** Suppose $X_n$ is a sequence of random vectors.
(i) Show $X_n \xrightarrow{P} 0$ if and only if $|X_n| \xrightarrow{P} 0$ (where the first zero refers to the zero vector and the second to the real number zero).
(ii) Show that convergence in probability of $X_n$ to $X$ is equivalent to convergence in probability of their components to the respective components of $X$.

**Problem 11.34** Assume $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, where $c$ is a constant. Show that $(X_n, Y_n) \xrightarrow{d} (X, c)$.

**Problem 11.35** Generalize Slutsky's Theorem (Theorem 11.3.2) to the case where $X_n$ is a vector, $A_n$ is a matrix, and $B_n$ is a vector.

**Problem 11.36** Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables. Write $X_i = X_i^+ - X_i^-$, where $X_i^+ = \max(X_i, 0)$. Suppose $X_i^-$ has a finite mean, but $X_i^+$ does not. Let $\bar{X}_n$ be the sample mean. Show $\bar{X}_n \xrightarrow{P} \infty$. *Hint:* For $B > 0$, let $Y_i = X_i$ if $X_i \le B$ and $Y_i = B$ otherwise; apply the Weak Law to $\bar{Y}_n$.

**Problem 11.37** (i) Let $K(P_0, P_1)$ be the Kullback–Leibler Information, defined in (11.21). Show that $K(P_0, P_1) \ge 0$ with equality iff $P_0 = P_1$.
(ii) Show the convergence (11.20) holds even when $K(P_0, P_1) = \infty$. *Hint:* Use Problem 11.36.

**Problem 11.38** As in Example 11.3.1, consider the problem of testing $P = P_0$ versus $P = P_1$ based on $n$ i.i.d. observations. The problem is an alternative way to show that a most powerful level $\alpha$ $(0 < \alpha < 1)$ test sequence has limiting power one. If

$P_0$ and $P_1$ are distinct, there exists $E$ such that $P_0(E) \neq P_1(E)$. Let $\hat{p}_n$ denote the proportion of observations in $E$ and construct a level-$\alpha$ test sequence based on $\hat{p}_n$ which has power tending to one.

**Problem 11.39** If $X_n$ is a sequence of real-valued random variables, prove that $X_n \to 0$ in $P_n$-probability if and only if $E_{P_n}[\min(|X_n|, 1)] \to 0$.

**Problem 11.40** (i) Prove Corollary 11.3.1.
(ii) Suppose $X_n \overset{d}{\to} X$ and $C_n \overset{P}{\to} \infty$. Show $P\{X_n \leq C_n\} \to 1$.

**Problem 11.41** In Example 11.3.2, show that $\beta_n(p_n) \to 1$ if $n^{1/2}(p_n - 1/2) \to \infty$ and $\beta_n(p_n) \to \alpha$ if $n^{1/2}(p_n - 1/2) \to 0$.

**Problem 11.42** In Example 11.3.4, let $I_n$ be the interval (11.23). Show that, for any $n$,

$$\inf_p P_p\{p \in \hat{I}_n\} = 0 \ .$$

*Hint:* Consider $p$ positive but small enough so that the chance that a sample of size $n$ results in 0 successes is nearly 1.

**Problem 11.43** Show how the interval (11.25) is obtained from (11.24).

**Problem 11.44** Prove Lemma 11.3.1

**Problem 11.45** Show that tightness of a sequence of random vectors in $\mathbb{R}^k$ is equivalent to each of the component variables being tight.

**Problem 11.46** Suppose $P_n$ is a sequence of probabilities and $X_n$ is a sequence of real-valued random variables; the distribution of $X_n$ under $P_n$ is denoted $\mathcal{L}(X_n|P_n)$. Prove that $\mathcal{L}(X_n|P_n)$ is tight if and only if $X_n/a_n \to 0$ in $P_n$-probability for every sequence $a_n \uparrow \infty$.

**Problem 11.47** Suppose $X_n \overset{d}{\to} N(\mu, \sigma^2)$. (i). Show that, for any sequence of numbers $c_n$, $P(X_n = c_n) \to 0$. (ii). If $c_n$ is any sequence such that $P(X_n > c_n) \to \alpha$, then $c_n \to \mu + \sigma z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1-\alpha$-quantile of $N(0, 1)$.

**Problem 11.48** Let $X_1, \cdots, X_n$ be i.i.d. normal with mean $\theta$ and variance 1. Suppose $\hat{\theta}_n$ is a location equivariant sequence of estimators such that, for every fixed $\theta$, $n^{1/2}(\hat{\theta}_n - \theta)$ converges in distribution to the standard normal distribution (if $\theta$ is true). Let $\bar{X}_n$ be the usual sample mean. Show that, if $\theta$ is fixed at the true value, then $n^{1/2}(\hat{\theta}_n - \bar{X}_n)$ tends to 0 in probability under $\theta$.

**Problem 11.49** Prove part (ii) of Theorem 11.3.4.

**Problem 11.50** Suppose $R$ is a real-valued function on $\mathbb{R}^k$ with $R(y) = o(|y|^p)$ as $|y| \to 0$, for some $p > 0$. If $Y_n$ is a sequence of random vectors satisfying $|Y_n| = o_P(1)$, then show $R(Y_n) = o_P(|Y_n|^p)$. *Hint:* Let $g(y) = R(y)/|y|^p$ with $g(0) = 0$ so that $g$ is continuous at 0; apply the Continuous Mapping Theorem.

**Problem 11.51** Use Problem 11.50 to prove (11.28).

**Problem 11.52** Assume $(U_i, V_i)$ is bivariate normal with correlation $\rho$. Let $\hat{\rho}_n$ denote the sample correlation given by (11.29). Verify the limit result (11.31).

**Problem 11.53** Consider the setting of Problem 6.21, where $(X_i, Y_i)$ are independent $N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$. The parameters $\mu_1, \ldots, \mu_n$ and $\sigma^2$ are all unknown. For testing $\sigma = 1$ against $\sigma > 1$, determine the limiting power of the UMPI level-$\alpha$ test against alternatives $1 + hn^{-1/2}$.

**Problem 11.54** (i) If $X_1, \ldots, X_n$ is a sample from a Poisson distribution with mean $E(X_i) = \lambda$, then $\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda})$ tends in law to $N(0, \frac{1}{4})$ as $n \to \infty$.
(ii) If $X$ has the binomial distribution $b(p, n)$, then $\sqrt{n}[\arcsin \sqrt{X/n} - \arcsin \sqrt{p}]$ tends in law to $N(0, \frac{1}{4})$ as $n \to \infty$.
*Note*. Certain refinements of variance stabilizing transformations are discussed by Anscombe (1948), Freeman and Tukey (1950), and Hotelling (1953). Transformations of data to achieve approximately a normal linear model are considered by Box and Cox (1964); for later developments stemming from this work see Bickel and Doksum (1981), Box and Cox (1982), and Hinkley and Runger (1984).

**Problem 11.55** Suppose $(X_1, \ldots, X_k)$ is multinomial based on $n$ trials and cell probabilities $(p_1, \ldots, p_k)$. Show that

$$\sqrt{n}\left[\sum_{j=1}^{k} \frac{X_j}{n} \log\left(\frac{X_j}{n}\right) - c\right]$$

converges in distribution to $F$, for some constant $c$ and distribution $F$. Identify $c$ and $F$.

**Problem 11.56** Suppose $X_{i,j}$ are independently distributed as $N(\mu_i, \sigma_i^2)$; $i = 1, \ldots, s$; $j = 1, \ldots, n_i$. Let $S_{n,i}^2 = \sum_j (X_{i,j} - \bar{X}_i)^2$, where $\bar{X}_i = n_i^{-1} \sum_j X_{i,j}$. Let $Z_{n,i} = \log[S_{n,i}^2/(n_i - 1)]$. Show that, as $n_i \to \infty$,

$$\sqrt{n_i - 1}[Z_{n,i} - \log(\sigma_i^2)] \xrightarrow{d} N(0, 2) .$$

Thus, for large $n_i$, the problem of testing equality of all the $\sigma_i$ can be approximately viewed as testing equality of means of normally distributed variables with known (possibly different) variances. Use Problem 7.12 to suggest a test.

**Problem 11.57** Let $X_1, \cdots, X_n$ be i.i.d. Poisson with mean $\lambda$. Consider estimating $g(\lambda) = e^{-\lambda}$ by the estimator $T_n = e^{-\bar{X}_n}$. Find an approximation to the bias of $T_n$; specifically, find a function $b(\lambda)$ satisfying

$$E_\lambda(T_n) = g(\lambda) + n^{-1}b(\lambda) + O(n^{-2})$$

as $n \to \infty$. Such an expression suggests a new estimator $T_n - n^{-1}b(\lambda)$, which has bias $O(n^{-2})$. But, $b(\lambda)$ is unknown. Show that the estimator $T_n - n^{-1}b(\bar{X}_n)$ has bias $O(n^{-2})$.

**Problem 11.58** Let $X_1, \ldots, X_n$ be a random sample from the Poisson distribution with unknown mean $\lambda$. The uniformly minimum variance unbiased estimator (UMVUE) of $\exp(-\lambda)$ is known to be $[(n-1)/n]^{T_n}$, where $T_n = \sum_{i=1}^n X_i$. Find the asymptotic distribution of the UMVUE (appropriately normalized). *Hint:* It may be easier to first find the asymptotic distribution of $\exp(-T_n/n)$.

**Problem 11.59** Let $X_{i,j}$, $1 \le i \le I$, $1 \le j \le n$ be independent with $X_{i,j}$ Poisson with mean $\lambda_i$. The problem is to test the null hypothesis that the $\lambda_i$ are all the same versus they are not all the same. Consider the test that rejects the null hypothesis iff

$$T \equiv \frac{n \sum_{i=1}^I (\bar{X}_i - \bar{X})^2}{\bar{X}}$$

is large, where $\bar{X}_i = \sum_j X_{i,j}/n$ and $\bar{X} = \sum_i \bar{X}_i/I$.
(i) How large should the critical values be so that, if the null hypothesis is correct, the probability of rejecting the null hypothesis tends (as $n \to \infty$ with $I$ fixed) to the nominal level $\alpha$.
(ii) Show that the test is pointwise consistent in power against any $(\lambda_1, \ldots, \lambda_I)$, as long as the $\lambda_i$ are not all equal.

**Problem 11.60** Assume $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$. Let $\hat{\sigma}_n^2$ be the maximum likelihood estimator of $\sigma^2$ given by $\hat{\sigma}_n^2 = \sum_{i=1}^n X_i^2/n$.
(i) Find the limiting distribution of $\sqrt{n}(\hat{\sigma}_n - \sigma)$.
(ii) For a constant $c$, let $T_{n,c} = c \sum_{i=1}^n |X_i|/n$. For what constant $c$ is $T_{n,c}$ a consistent estimator of $\sigma$?
(iii) Determine the limiting distribution of $\sqrt{n}(T_{n,c} - \sigma)$ with $c$ chosen as your consistent estimator.
(iv) Determine the limiting distribution of $\sqrt{n}\log(\hat{\sigma}_n/T_{n,c})$ (again with $c$ chosen from (ii) above).

**Problem 11.61** Suppose $X_1, \ldots, X_I$ are independent and binomially distributed, with $X_i \sim b(n_i, p_i)$; that is, $X_i$ is the number of successes in $n_i$ Bernoulli trials. Suppose that $p_i$ satisfies

$$\log[p_i/(1-p_i)] = \theta d_i$$

for known constants $d_i$, which implies

$$p_i = \frac{e^{d_i\theta}}{1 + e^{d_i\theta}}$$

(Think of $d_i$ as the dose given to $n_i$ subjects, and you observe $X_i$ deaths at the dosage level $d_i$.) Both $d_i$ and $n_i$ are known.

(i) For testing the null hypothesis $\theta = 0$ against $\theta = 1$, find the form of the most powerful level-$\alpha$ test and show that it rejects for large values of a test statistic $T$.
(ii) If the null hypothesis is true and the sample sizes $n_i$ are moderately large, what is the approximate distribution of $T$?
(iii) If $I = 5$, $d_i = i$ and $n_i = 100$, approximate the $p$-value of your test if you observe $X_1 = 40$, $X_2 = 51$, $X_3 = 64$, $X_4 = 73$ and $X_5 = 80$.

## *Section 11.4*

**Problem 11.62** Prove the Glivenko–Cantelli Theorem. *Hint:* Use the Strong Law of Large Numbers and the monotonicity of $F$.

**Problem 11.63** Let $X_1, \ldots, X_n$ be i.i.d. $P$ on $S$. Suppose $S$ is countable and let $\mathcal{E}$ be the collection of *all* subsets of $S$. Let $\hat{P}_n$ be the *empirical measure*; that is, for any subset $E$ of $\mathcal{E}$, $\hat{P}_n(E)$ is the proportion of observations $X_i$ that fall in $E$. Prove, with probability one,

$$\sup_{E \in \mathcal{E}} |\hat{P}_n(E) - P(E)| \to 0 .$$

**Problem 11.64** Suppose $X_n$ is a tight sequence and $Y_n \xrightarrow{P} 0$. Show that $X_n Y_n \xrightarrow{P} 0$. If it is assumed $Y_n \to 0$ almost surely, can you conclude $X_n Y_n \to 0$ almost surely?

**Problem 11.65** Suppose $X_n$ is a sequence of real-valued random variables.
(i) Assume $X_n$ is Cauchy in probability; that is, for all $\epsilon > 0$,

$$\lim_{\min(m,n)\to\infty} P\{|X_n - X_m| > \epsilon\} \to 0 .$$

Then, show there exists a random variable $X$ such that $X_n \xrightarrow{P} X$, in which case we may write $X = \lim_{n\to\infty} X_n$.
(ii) Assume $X_n$ satisfies $E(|X_n|^p) < \infty$. Also, assume $X_n$ is Cauchy in $L_p$; that is,

$$\lim_{\min(m,n)\to\infty} E(|X_n - X_m|^p) \to 0 .$$

Then, show there exist a random variable $X$ such that $E(|X_n - X|^p) \to 0$ and $E(|X|^p) < \infty$.

**Problem 11.66** For a c.d.f. $F$, define the quantile transformation $Q$ by

$$Q(u) = \inf\{t : F(t) \geq u\} .$$

(i) Show the event $\{F(t) \geq u\}$ is the same as $\{Q(u) \leq t\}$.
(ii) If $U$ is uniformly distributed on (0, 1), show the distribution of $Q(U)$ is $F$.

**Problem 11.67** Assume $X_n$ has c.d.f. $F_n$. Fix $\alpha \in (0, 1)$.
(i) If $X_n$ is tight, show that $F_n^{-1}(1 - \alpha)$ is uniformly bounded.
(ii) If $X_n \xrightarrow{P} c$, show that $F_n^{-1}(1 - \alpha) \to c$.

**Problem 11.68** Let $U_1, \dots, U_n$ be i.i.d. with c.d.f. $G(u) = u$ and let $\hat{G}_n$ denote the empirical c.d.f. of $U_1, \dots, U_n$. Define

$$B_n(u) = n^{1/2}[\hat{G}_n(u) - u] \ .$$

(Note that $B_n(\cdot)$ is a random function, called the *uniform empirical process*).
(i) Show that the distribution of the Kolmogorov–Smirnov test statistic $n^{1/2} d_K(\hat{G}_n, G)$ under $G$ is that of $\sup_u |B_n(u)|$.
(ii) Suppose $X_1, \dots, X_n$ are i.i.d. $F$ (not necessarily continuous), and let $\hat{F}_n$ denote the empirical c.d.f. of $X_1, \dots, X_n$. Show that the distribution of the Kolmogorov–Smirnov test statistic $n^{1/2} d_K(\hat{F}_n, F)$ under $F$ is that of $\sup_t |B_n(F(t))|$, where $B_n$ is defined in (i). Deduce that this distribution does not depend on $F$ when $F$ is continuous.

**Problem 11.69** Consider the uniform confidence band $R_{n,1-\alpha}$ for $F$ given by (11.36). Let $\mathbf{F}$ be the set of all distributions on $\mathbb{R}$. Show,

$$\inf_{F \in \mathbf{F}} P_F\{F \in R_{n,1-\alpha}\} \geq 1 - \alpha \ .$$

**Problem 11.70** Show how Theorem 11.4.3 implies Theorem 11.4.2. *Hint:* Use the Borel–Cantelli Lemma; see Billingsley (1995, Theorem 4.3).

**Problem 11.71** (i) If $X_1, \dots, X_n$ are i.i.d. with c.d.f. $F$ and empirical distribution $\hat{F}_n$, use Theorem 11.4.3 to show that $n^{1/2} \sup |\hat{F}_n(t) - F(t)|$ is a tight sequence.
(ii) Let $F_n$ be any sequence of distributions, and let $\hat{F}_n$ be the empirical distribution based on a sample of size $n$ from $F_n$. Show that $n^{1/2} \sup |\hat{F}_n(t) - F_n(t)|$ is a tight sequence.

**Problem 11.72** Show that $X_n \to X$ in probability is equivalent to the statement that, for any subsequence $X_{n_j}$, there exists a further subsequence $X_{n_{j_k}}$ such that $X_{n_{j_k}} \to X$ with probability one.

**Problem 11.73** (i) Suppose random variables $X_n$, $Y_n$ and a random vector $W_n$ are such that, given $W_n$, $X_n$ and $Y_n$ are conditionally independent. Assume, for nonnegative constants $\sigma_X$ and $\sigma_Y$, and for all $z$,

$$P\{X_n \leq z | W_n\} \xrightarrow{P} \Phi(z/\sigma_X\}$$

and

$$P\{Y_n \leq z | W_n\} \xrightarrow{P} \Phi(z/\sigma_Y\} \ .$$

Show that

$$P\{X_n + Y_n \le z | W_n\} \xrightarrow{P} \Phi(z/\sqrt{\sigma_X^2 + \sigma_Y^2}) \ .$$

How do the unconditional distributions behave?
(ii) Suppose $\hat{F}_n$ and $\hat{G}_n$ are (random) distributions, and assume $F = N(0, \sigma_F^2)$ and $G = N(0, \sigma_G^2)$ are nonrandom. Let $\rho$ be any metric metrizing weak convergence, such as the Lévy metric. If

$$\rho(\hat{F}_n, F) \xrightarrow{P} 0$$

and

$$\rho(\hat{G}_N, G) \xrightarrow{P} 0 \ ,$$

then show

$$\rho(\hat{F}_n * \hat{G}_n, F * G) \xrightarrow{P} 0 \ ,$$

where $F * G$ denotes the convolution between distributions $F$ and $G$.

**Problem 11.74** (i) Show that if $\{X_n\}$ is uniformly integrable, then $\{X_n\}$ is asymptotically uniformly integrable, but the converse is false.
(ii) Show that a sufficient condition for $\{X_n\}$ to be uniformly integrable is, for some $\delta > 0$, $\sup_n E(|X_n|^{1+\delta}) < \infty$.

**Problem 11.75** If $X_n \xrightarrow{P} 0$ and

$$\sup_n E[|X_n|^{1+\delta}] < \infty \quad \text{for some } \delta > 0 \ , \tag{11.42}$$

then show $E[|X_n|] \to 0$. (More generally, if the $X_n$ are *uniformly integrable* in the sense $\sup_n E[|X_n| I\{|X_n| > t\}] \to 0$ as $t \to \infty$, then $E[|X_n|] \to 0$. A converse is given in Dudley (1989), p. 279.)

**Problem 11.76** (i) Show that $\{X_n\}$ is uniformly integrable if and only if $\sup_n E|X_n| < \infty$ and

$$\sup_n E[|X_n| I_A\} = \int_A |X_n(\omega)| dP(\omega) \to 0$$

as $P\{A\} \to 0$.
(ii) Suppose $X_1, \ldots, X_n$ are i.i.d. with finite mean $\mu$. Show that $\bar{X}_n$ is uniformly integrable and hence $E|\bar{X}_n - \mu| \to 0$. (The fact that $\bar{X}_n$ is uniformly integrable holds if the $X_i$ are just identically distributed with finite mean.)

**Problem 11.77** If $X_n \xrightarrow{d} X$ and $\{X_n\}$ is asymptotically uniformly integrable, show that for any $0 < p < 1$, $E(X_n^p) \to E(X^p)$.

**Problem 11.78** Assume $X_1, \ldots, X_n$ are i.i.d. with $E(|X_i|^p] < \infty$. Then, show that

$$n^{-\frac{1}{p}} \max_{1 \le i \le n} |X_i| \xrightarrow{P} 0 \, .$$

**Problem 11.79** (i) Suppose $X_n \xrightarrow{d} X$ and $Var(X_n) \to Var(X) < \infty$. Show $E(X_n) \to E(X)$.

(ii) Suppose $(X_n, Y_n) \xrightarrow{d} (X, Y)$ in the plane, with $Var(X_n) \to Var(X) < \infty$ and $Var(Y_n) \to Var(Y) < \infty$. Show that $Cov(X_n, Y_n) \to Cov(X, Y)$.

## 11.6 Notes

The convergence concepts in this chapter are classical and can be found in most graduate probability texts such as Billingsley (1995) or Dudley (1989). The Central Limit Theory for Bernoulli trials dates back to de Moivre (1733) and for more general distributions to Laplace (1812). Their treatment was probabilistic and did not involve problems in inference. Normal experiments were first treated in Gauss (1809). Further history is provided in Stigler (1986) and Hald (1990, 1998).