Approximation Theorems of Mathematical Statistics

ROBERT J. SERFLING
The Johns Hopkins University

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

This Page Intentionally Left Blank

Approximation Theorems of Mathematical Statistics

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Approximation Theorems of Mathematical Statistics

ROBERT J. SERFLING
The Johns Hopkins University

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

This text is printed on acid-free paper. ⊗

Paperback edition published 2002.

Published simultaneously in Canada.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging in Publication Data is available.

ISBN 0-471-21927-4

To my parents and to
the memory of my wife's parents

This Page Intentionally Left Blank

Preface

This book covers a broad range of limit theorems useful in mathematical statistics, along with methods of proof and techniques of application. The manipulation of "probability" theorems to obtain "statistical" theorems is emphasized. It is hoped that, besides a knowledge of these basic statistical theorems, an appreciation on the instrumental role of probability theory and a perspective on practical needs for its further development may be gained.

A one-semester course each on probability theory and mathematical statistics at the beginning graduate level is presupposed. However, highly polished expertise is not necessary, the treatment here being self-contained at an elementary level. The content is readily accessible to students in statistics, general mathematics, operations research, and selected engineering fields.

Chapter 1 lays out a variety of tools and foundations basic to asymptotic theory in statistics as treated in this book. Foremost are: modes of convergence of a sequence of random variables (convergence in distribution, convergence in probability, convergence almost surely, and convergence in the rth mean); probability limit laws (the law of large numbers, the central limit theorem, and related results).

Chapter 2 deals systematically with the usual statistics computed from a sample: the sample distribution function, the sample moments, the sample quantiles, the order statistics, and cell frequency vectors. Properties such as asymptotic normality and almost sure convergence are derived. Also, deeper insights are pursued, including R. R. Bahadur's fruitful almost sure representations for sample quantiles and order statistics. Building on the results of Chapter 2, Chapter 3 treats the asymptotics of statistics concocted as transformations of vectors of more basic statistics. Typical examples are the sample coefficient of variation and the chi-squared statistic. Taylor series approximations play a key role in the methodology.

The next six chapters deal with important special classes of statistics. Chapter 4 concerns statistics arising in classical parametric inference and contingency table analysis. These include maximum likelihood estimates,

likelihood ratio tests, minimum chi-square methods, and other asymptotically efficient procedures.

Chapter 5 is devoted to the sweeping class of W. Hoeffding's U-statistics, which elegantly and usefully generalize the notion of a sample mean. Basic convergence theorems, probability inequalities, and structural properties are derived. Introduced and applied here is the important "projection" method, for approximation of a statistic of arbitrary form by a simple sum of independent random variables.

Chapter 6 treats the class of R. von Mises' "differentiable statistical functions," statistics that are formulated as functionals of the sample distribution function. By differentiation of such a functional in the sense of the Gâteaux derivative, a reduction to an approximating statistic of simpler structure (essentially a U-statistic) may be developed, leading in a quite mechanical way to the relevant convergence properties of the statistical function. This powerful approach is broadly applicable, as most statistics of interest may be expressed either exactly or approximately as a "statistical function."

Chapters 7, 8, and 9 treat statistics obtained as solutions of equations ("M-estimates"), linear functions of order statistics ("L-estimates"), and rank statistics ("R-estimates"), respectively, three classes important in robust parametric inference and in nonparametric inference. Various methods, including the projection method introduced in Chapter 5 and the differential approach of Chapter 6, are utilized in developing the asymptotic properties of members of these classes.

Chapter 10 presents a survey of approaches toward asymptotic relative efficiency of statistical test procedures, with special emphasis on the contributions of E. J. G. Pitman, H. Chernoff, R. R. Bahadur, and W. Hoeffding.

To get to the end of the book in a one-semester course, some time-consuming material may be skipped without loss of continuity. For example, Sections 1.4, 1.11, 2.8, 3.6, and 4.3, and the proofs of Theorems 2.3.3C and 9.2.6A, B, C, may be so omitted.

This book evolved in conjunction with teaching such a course at The Florida State University in the Department of Statistics, chaired by R. A. Bradley. I am thankful for the stimulating professional environment conducive to this activity. Very special thanks are due D. D. Boos for collaboration on portions of Chapters 6, 7, and 8 and for many useful suggestions overall. I also thank J. Lynch, W. Pirie, R. Randles, I. R. Savage, and J. Sethuraman for many helpful comments. To the students who have taken this course with me, I acknowledge warmly that each has contributed a constructive impact on the development of this book. The support of the Office of Naval Research, which has sponsored part of the research in Chapters 5, 6, 7, 8, and 9 is acknowledged with appreciation. Also, I thank Mrs. Kathy

Strickland for excellent typing of the manuscript. Finally, most important of all, I express deep gratitude to my wife, Jackie, for encouragement without which this book would not have been completed.

ROBERT J. SERFLING

Baltimore, Maryland
September 1980

This Page Intentionally Left Blank

Contents

Approximation Theorems of Mathematical Statistics

This Page Intentionally Left Blank

CHAPTER 1

Preliminary Tools and Foundations

This chapter lays out tools and foundations basic to asymptotic theory in statistics as treated in this book. It is intended to reinforce previous knowledge as well as perhaps to fill gaps. As for actual proficiency, that may be gained in later chapters through the process of implementation of the material.

Of particular importance, Sections 1.2–1.7 treat notions of convergence of a sequence of random variables, Sections 1.8–1.11 present key probability limit theorems underlying the statistical limit theorems to be derived, Section 1.12 concerns differentials and Taylor series, and Section 1.15 introduces concepts of asymptotics of interest in the context of statistical inference procedures.

1.1 PRELIMINARY NOTATION AND DEFINITIONS

1.1.1 Greatest Integer Part

For x real, $[x]$ denotes the greatest integer less than or equal to x.

1.1.2 $O(\cdot)$, $o(\cdot)$, and $\sim$

These symbols are called "big oh," "little oh," and "twiddle," respectively. They denote ways of comparing the magnitudes of two functions $u(x)$ and $v(x)$ as the argument x tends to a limit L (not necessarily finite). The notation $u(x) = O(v(x))$, $x \to L$, denotes that $|u(x)/v(x)|$ remains bounded as $x \to L$. The notation $u(x) = o(v(x))$, $x \to L$, stands for

$$\lim_{x \to L} \frac{u(x)}{v(x)} = 0,$$

and the notation $u(x) \sim v(x)$, $x \to L$, stands for

$$\lim_{x \to L} \frac{u(x)}{v(x)} = 1.$$

Probabilistic versions of these "order of magnitude" relations are given in **1.2.6**, after introduction of some convergence notions.

Example. Consider the function

$$f(n) = 1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right).$$

Obviously, $f(n) \to 0$ as $n \to \infty$. But we can say more. Check that

$$f(n) = \frac{3}{n} + O(n^{-2}), \, n \to \infty,$$

$$= \frac{3}{n} + o(n^{-1}), \, n \to \infty,$$

$$\sim \frac{3}{n}, \, n \to \infty. \quad \blacksquare$$

1.1.3 Probability Space, Random Variables, Random Vectors

In our discussions there will usually be (sometimes only implicitly) an underlying *probability space* $(\Omega, \mathscr{A}, P)$, where Ω is a set of points, $\mathscr{A}$ is a σ-field of subsets of Ω, and P is a probability distribution or measure defined on the elements of $\mathscr{A}$. A *random variable* $X(\omega)$ is a transformation of Ω into the real line R such that images $X^{-1}(B)$ of Borel sets B are elements of $\mathscr{A}$. A collection of random variables $X_1(\omega), X_2(\omega), \ldots$ on a given pair $(\Omega, \mathscr{A})$ will typically be denoted simply by $X_1, X_2, \ldots$. A *random vector* is a k-tuple $\mathbf{X} = (X_1, \ldots, X_k)$ of random variables defined on a given pair $(\Omega, \mathscr{A})$.

1.1.4 Distributions, Laws, Expectations, Quantiles

Associated with a random vector $\mathbf{X} = (X_1, \ldots, X_k)$ on $(\Omega, \mathscr{A}, P)$ is a right-continuous *distribution function* defined on R^k by

$$F_{X_1, \ldots, X_k}(t_1, \ldots, t_k) = P(\{\omega: X_1(\omega) \le t_1, \ldots, X_k(\omega) \le t_k\})$$

for all $\mathbf{t} = (t_1, \ldots, t_k) \in R^k$. This is also known as the *probability law* of $\mathbf{X}$. (There is also a left-continuous version.) Two random vectors $\mathbf{X}$ and $\mathbf{Y}$, defined on possibly different probability spaces, "have the same law" if their distribution functions are the same, and this is denoted by $\mathscr{L}(\mathbf{X}) = \mathscr{L}(\mathbf{Y})$, or $F_{\mathbf{X}} = F_{\mathbf{Y}}$.

By *expectation* of a random variable X is meant the Lebesgue–Stieltjes integral of $X(\omega)$ with respect to the measure P. Commonly used notations for this expectation are $E\{X\}$, EX, $\int_\Omega X(\omega)dP(\omega)$, $\int_\Omega X(\omega)P(d\omega)$, $\int X\,dP$, $\int X$, $\int_{-\infty}^{\infty} t\,dF_X(t)$, and $\int t\,dF_X$. All denote the same quantity. Expectation may also be represented as a Riemann–Stieltjes integral (see Cramér (1946), Sections 7.5 and 9.4). The expectation $E\{X\}$ is also called the *mean* of the random variable X. For a random *vector* $\mathbf{X} = (X_1, \ldots, X_k)$, the mean is defined as $E\{\mathbf{X}\} = (E\{X_1\}, \ldots, E\{X_k\})$.

Some important characteristics of random variables may be represented conveniently in terms of expectations, provided that the relevant integrals exist. For example, the *variance* of X is given by $E\{(X - E\{X\})^2\}$, denoted $\mathrm{Var}\{X\}$. More generally, the *covariance* of two random variables X and Y is given by $E\{(X - E\{X\})(Y - E\{Y\})\}$, denoted $\mathrm{Cov}\{X, Y\}$. (Note that $\mathrm{Cov}\{X, X\} = \mathrm{Var}\{X\}$.) Of course, such an expectation may also be represented as a Riemann–Stieltjes integral,

$$\mathrm{Cov}\{X, Y\} = \iint (x - E\{X\})(y - E\{Y\})dF_{X,Y}(x, y).$$

For a random vector $\mathbf{X} = (X_1, \ldots, X_k)$, the *covariance matrix* is given by $\mathbf{\Sigma} = (\sigma_{ij})_{k \times k}$, where $\sigma_{ij} = \mathrm{Cov}\{X_i, X_j\}$.

For any univariate distribution function F, and for $0 < p < 1$, the quantity

$$F^{-1}(p) = \inf\{x: F(x) \geq p\}$$

is called the pth *quantile* or *fractile* of F. It is also denoted ξ_p. In particular, $\xi_{1/2} = F^{-1}(\frac{1}{2})$ is called the *median* of F.

The function $F^{-1}(t)$, $0 < t < 1$, is called the *inverse* function of F. The following proposition, giving useful properties of F and F^{-1}, is easily checked (Problem 1.P.1).

Lemma. *Let* F *be a distribution function. The function* $\mathrm{F}^{-1}(\mathrm{t})$, $0 < \mathrm{t} < 1$, *is nondecreasing and left-continuous, and satisfies*

(i) $\mathrm{F}^{-1}(\mathrm{F}(\mathrm{x})) \leq \mathrm{x}$, $-\infty < \mathrm{x} < \infty$,

and

(ii) $\mathrm{F}(\mathrm{F}^{-1}(\mathrm{t})) \geq \mathrm{t}$, $0 < \mathrm{t} < 1$.

Hence

(iii) $\mathrm{F}(\mathrm{x}) \geq \mathrm{t}$ *if and only if* $\mathrm{x} \geq \mathrm{F}^{-1}(\mathrm{t})$.

A further useful lemma, concerning the inverse functions of a weakly convergent sequence of distributions, is given in **1.5.6**.

1.1.5 $N(\mu, \sigma^2)$, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The *normal* distribution with mean μ and variance $\sigma^2 > 0$ corresponds to the distribution function

$$F(x) = \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt, \qquad -\infty < x < \infty.$$

The notation $N(\mu, \sigma^2)$ will be used to denote either this distribution or a random variable having this distribution—whichever is indicated by the context. The special distribution function $N(0, 1)$ is known as the *standard normal* and is often denoted by Φ. In the case $\sigma^2 = 0$, $N(\mu, \sigma^2)$ will denote the distribution *degenerate* at μ, that is, the distribution

$$F(x) = \begin{cases} 0, & x < \mu, \\ 1, & x \geq \mu. \end{cases}$$

A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ has the *k-variate normal* distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{k \times k}$ if, for every nonnull vector $\mathbf{a} = (a_1, \ldots, a_k)$, the random variable $\mathbf{aX}'$ is $N(\mathbf{a\mu}', \mathbf{a\Sigma a}')$, that is, $\mathbf{aX}' = \sum_{i=1}^{k} a_i X_i$ has the normal distribution with mean $\mathbf{a\mu}' = \sum_1^k a_i \mu_i$ and variance $\mathbf{a\Sigma a}' = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j \sigma_{ij}$. The notation $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ will denote either this multivariate distribution or a random vector having this distribution.

The components X_i of a multivariate normal vector are seen to have (univariate) normal distributions. However, the converse does not hold. Random variables $X_1, \ldots, X_k$ may each be normal, yet possess a joint distribution which is *not* multivariate normal. Examples are discussed in Ferguson (1967), Section 3.2.

1.1.6 Chi-squared Distributions

Let $\mathbf{Z}$ be k-variate $N(\boldsymbol{\mu}, \mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix of order k. For the case $\boldsymbol{\mu} = \mathbf{0}$, the distribution of $\mathbf{ZZ}' = \sum_1^k Z_i^2$ is called the *chi-squared with k degrees of freedom*. For the case $\boldsymbol{\mu} \neq \mathbf{0}$, the distribution is called *noncentral chi-squared with k degrees of freedom and noncentrality parameter* $\lambda = \boldsymbol{\mu\mu}'$. The notation $\chi_k^2(\lambda)$ encompasses both cases and may denote either the random variable or the distribution. We also denote $\chi_k^2(0)$ simply by χ_k^2.

1.1.7 Characteristic Functions

The *characteristic function* of a random k-vector $\mathbf{X}$ is defined as

$$\phi_{\mathbf{X}(\mathbf{t})} = E\{e^{i\mathbf{t}\mathbf{X}'}\} = \int \cdots \int e^{i\mathbf{t}\mathbf{x}'} \, dF_{\mathbf{X}(\mathbf{x})}, \qquad \mathbf{t} \in R^k.$$

In particular, the characteristic function of $N(0, 1)$ is $\exp(-\frac{1}{2}t^2)$. See Lukacs (1970) for a full treatment of characteristic functions.

1.1.8 Absolutely Continuous Distribution Functions

An *absolutely continuous* distribution function F is one which satisfies

$$F(x) = \int_{-\infty}^{x} F'(t)dt, \qquad -\infty < x < \infty.$$

That is, F may be represented as the indefinite integral of its derivative. In this case, any function f such that $F(x) = \int_{-\infty}^{x} f(t)dt$, all x, is called a *density* for F. Any such density must agree with F' except possibly on a Lebesgue-null set. Further, if f is continuous at x_0, then $f(x_0) = F'(x_0)$ must hold. This latter may be seen by elementary arguments. For detailed discussion, see Natanson (1961), Chapter IX.

1.1.9 I.I.D.

With reference to a sequence $\{\mathbf{X}_i\}$ of random vectors, the abbreviation *I.I.D.* will stand for "independent and identically distributed."

1.1.10 Indicator Functions

For any set S, the associated *indicator function* is

$$I_S(x) = \begin{cases} 1, & x \in S, \\ 0, & x \notin S. \end{cases}$$

For convenience, the alternate notation $I(S)$ will sometimes be used for I_S, when the argument x is suppressed.

1.1.11 Binomial (n, p)

The *binomial* distribution with parameters n and p, where n is a positive integer and $0 \leq p \leq 1$, corresponds to the probability mass function

$$p(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

The notation $B(n, p)$ will denote either this distribution or a random variable having this distribution. As is well known, $B(n, p)$ is the distribution of the number of successes in a series of n independent trials each having success probability p.

1.1.12 Uniform (a, b)

The *uniform* distribution on the interval $[a, b]$, denoted $U(a, b)$, corresponds to the density function $f(x) = 1/(b-a)$, $a \leq x \leq b$, and $=0$, otherwise.

1.2 MODES OF CONVERGENCE OF A SEQUENCE OF RANDOM VARIABLES

Two forms of approximation are of central importance in statistical applications. In one form, a given random variable is approximated by another random variable. In the other, a given distribution function is approximated by another distribution function. Concerning the first case, three modes of convergence for a sequence of random variables are introduced in **1.2.1**, **1.2.2**, and **1.2.3**. These modes apply also to the second type of approximation, along with a fourth distinctive mode introduced in **1.2.4**. Using certain of these convergence notions, stochastic versions of the $O(\cdot)$, $o(\cdot)$ relations in **1.1.2** are introduced in **1.2.5**. A brief illustration of ideas is provided in **1.2.6**.

1.2.1 Convergence in Probability

Let $X_1, X_2, \ldots$ and X be random variables on a probability space $(\Omega, \mathcal{A}, P)$. We say that X_n *converges in probability* to X if

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1, \qquad \text{every} \quad \varepsilon > 0.$$

This is written $X_n \xrightarrow{p} X$, $n \to \infty$, or $p\text{-}\lim_{n\to\infty} X_n = X$. Examples are in **1.2.6**, Section 1.8, and later chapters. Extension to the case of $X_1, X_2, \ldots$ and X random elements of a metric space is straightforward, by replacing $|X_n - X|$ by the relevant metric (see Billingsley (1968)). In particular, for random k-vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots$ and $\mathbf{X}$, we shall say that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if $\|\mathbf{X}_n - \mathbf{X}\| \xrightarrow{p} 0$ in the above sense, where $\|\mathbf{z}\| = (\sum_{i=1}^{k} z_i^2)^{1/2}$ for $\mathbf{z} \in R^k$. It then follows (Problem 1.P.2) that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if and only if the corresponding component-wise convergences hold.

1.2.2 Convergence with Probability 1

Consider random variables $X_1, X_2, \ldots$ and X on $(\Omega, \mathcal{A}, P)$. We say that X_n *converges with probability 1* (or *strongly, almost surely, almost everywhere*, etc.) to X if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

This is written $X_n \xrightarrow{wp1} X$, $n \to \infty$, or $p1\text{-}\lim_{n\to\infty} X_n = X$. Examples are in **1.2.6**, Section 1.9, and later chapters. Extension to more general random elements is straightforward.

An equivalent condition for convergence $wp1$ is

$$\lim_{n\to\infty} P(|X_m - X| < \varepsilon, \text{ all } m \geq n) = 1, \qquad \text{each} \quad \varepsilon > 0.$$

This facilitates comparison with convergence in probability. The equivalence is proved by simple set-theoretic arguments (Halmos (1950), Section 22), as follows. First check that

$$(*)\quad \left\{\omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = \bigcap_{\varepsilon>0} \bigcup_{n=1}^{\infty} \{\omega: |X_m(\omega) - X(\omega)| < \varepsilon, \text{ all } m \geq n\},$$

whence

$$(**)\quad \left\{\omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = \lim_{\varepsilon\to 0} \lim_{n\to\infty} \{\omega: |X_m(\omega) - X(\omega)| < \varepsilon, \text{ all } m \geq n\}.$$

By the continuity theorem for probability functions (Appendix), (**) implies

$$P(X_n \to X) = \lim_{\varepsilon\to 0} \lim_{n\to\infty} P(|X_m - X| < \varepsilon, \text{all } m \geq n),$$

which immediately yields one part of the equivalence. Likewise, (*) implies, for any $\varepsilon > 0$,

$$P(X_n \to X) \leq \lim_{n\to\infty} P(|X_m - X| < \varepsilon, \text{ all } m \geq n),$$

yielding the other part.

The relation (*) serves also to establish that the set $\{\omega: X_n(\omega) \to X(\omega)\}$ truly belongs to $\mathscr{A}$, as is necessary for "convergence *wp*1" to be well defined.

A somewhat stronger version of this mode of convergence will be noted in **1.3.4.**

1.2.3 Convergence in *r*th Mean

Consider random variables $X_1, X_2, \ldots$ and X on $(\Omega, \mathscr{A}, P)$. For $r > 0$, we say that X_n *converges in* rth *mean* to X if

$$\lim_{n\to\infty} E|X_n - X|^r = 0.$$

This is written $X_n \xrightarrow{r\text{th}} X$ or $L_r\text{-}\lim_{n\to\infty} X_n = X$. The higher the value of r, the more stringent the condition, for an application of Jensen's inequality (Appendix) immediately yields

$$X_n \xrightarrow{r\text{th}} X \Rightarrow X_n \xrightarrow{s\text{th}} X, 0 < s < r.$$

Given $(\Omega, \mathscr{A}, P)$ and $r > 0$, denote by $L_r(\Omega, \mathscr{A}, P)$ the space of random variables Y such that $E|Y|^r < \infty$. The usual metric in L_r is given by $d(Y, Z) = \|Y - Z\|_r$, where

$$\|Y\|_r = \begin{cases} E|Y|^r, & 0 < r < 1, \\ [E|Y|^r]^{1/r}, & r \geq 1. \end{cases}$$

Thus convergence in the rth mean may be interpreted as convergence in the L_r metric, in the case of random variables $X_1, X_2, \ldots$ and X belonging to L_r.

1.2.4 Convergence in Distribution

Consider distribution functions $F_1(\cdot), F_2(\cdot), \ldots$ and $F(\cdot)$. Let $X_1, X_2, \ldots$ and X denote random variables (not necessarily on a common probability space) having these distributions, respectively. We say that X_n *converges in distribution* (or *in law*) to X if

$$\lim_{n\to\infty} F_n(t) = F(t), \text{ each continuity point } t \text{ of } F.$$

This is written $X_n \xrightarrow{d} X$, or $d\text{-}\lim_{n\to\infty} X_n = X$. A detailed examination of this mode of convergence is provided in Section 1.5. Examples are in **1.2.6**, Section 1.9, and later chapters.

The reader should figure out why this definition would *not* afford a satisfactory notion of approximation of a given distribution function by other ones if the convergence were required to hold for *all* t.

In as much as the definition of $X_n \xrightarrow{d} X$ is formulated wholly in terms of the corresponding distribution functions F_n and F, it is sometimes convenient to use the more direct notation "$F_n \Rightarrow F$" and the alternate terminology "F_n *converges weakly to* F." However, as in this book the discussions will tend to refer directly to various random variables under consideration, the notation $X_n \xrightarrow{d} X$ will be quite useful also.

Remark. The convergences $\xrightarrow{p}$, $\xrightarrow{wp1}$, and $\xrightarrow{r\text{th}}$ each represent a sense in which, for n sufficiently large, $X_n(\omega)$ and $X(\omega)$ approximate each other as *functions of* ω, $\omega \in \Omega$. This means that the *distributions* of X_n and X cannot be too dissimilar, whereby approximation in distribution should follow. On the other hand, the convergence $\xrightarrow{d}$ depends *only* on the distribution functions involved and does not necessitate that the relevant X_n and X approximate each other as functions of ω. In fact, X_n and X need not be defined on the same probability space. Section 1.3 deals formally with these interrelationships. ■

1.2.5 Stochastic $O(\cdot)$ and $o(\cdot)$

A sequence of random variables $\{X_n\}$, with respective distribution functions $\{F_n\}$, is said to be *bounded in probability* if for every $\varepsilon > 0$ there exist M_ε and N_ε such that

$$F_n(M_\varepsilon) - F_n(-M_\varepsilon) > 1 - \varepsilon \qquad \text{all} \quad n > N_\varepsilon.$$

The notation $X_n = O_p(1)$ will be used. It is readily seen that $X_n \xrightarrow{d} X \Rightarrow X_n = O_p(1)$ (Problem 1.P.3).

More generally, for two sequences of random variables $\{U_n\}$ and $\{V_n\}$, the notation $U_n = O_p(V_n)$ denotes that the sequence $\{U_n/V_n\}$ is $O_p(1)$. Further, the notation $U_n = o_p(V_n)$ denotes that $U_n/V_n \xrightarrow{p} 0$. Verify (Problem 1.P.4) that $U_n = o_p(V_n) \Rightarrow U_n = O_p(V_n)$.

1.2.6 Example: Proportion of Successes in a Series of Trials

Consider an infinite series of independent trials each having the outcome "success" with probability p. (The underlying probability space would be based on the set Ω of all infinite sequences ω of outcomes of such a series of trials.) Let X_n denote the proportion of successes in the first n trials. Then

(i) $X_n \xrightarrow{p} p$;

(ii) $X_n \xrightarrow{wp1} p$;

(iii) $\dfrac{\sqrt{n}(X_n - p)}{[p(1-p)]^{1/2}} \xrightarrow{d} N(0, 1)$;

(iv) $\dfrac{\sqrt{n}(X_n - p)}{(\log \log n)^{1/2}} \xrightarrow{p} 0$;

(v) $\dfrac{\sqrt{n}(X_n - p)}{(\log \log n)^{1/2}} \xrightarrow{wp1} 0$;

(vi) $X_n \xrightarrow{2nd} p$.

Is it true that

(vii) $\dfrac{\sqrt{n}(X_n - p)}{(\log \log n)^{1/2}} \xrightarrow{2nd} 0$?

Justification and answers regarding (i)–(v) await material to be covered in Sections 1.8–1.10. Items (vi) and (vii) may be resolved at once, however, simply by computing variances (Problem 1.P.5).

1.3 RELATIONSHIPS AMONG THE MODES OF CONVERGENCE

For the four modes of convergence introduced in Section 1.2, we examine here the key relationships as given by direct implications (**1.3.1–1.3.3**), partial converses (**1.3.4–1.3.7**), and various counter-examples (**1.3.8**). The question of convergence of moments, which is related to the topic of convergence in rth mean, is treated in Section 1.4.

1.3.1 Convergence *wp*1 Implies Convergence in Probability

Theorem. *If* $X_n \xrightarrow{wp1} X$, *then* $X_n \xrightarrow{p} X$.

This is an obvious consequence of the equivalence noted in **1.2.2.** Incidentally, the proposition is not true in general for *all* measures (e.g., see Halmos (1950)).

1.3.2 Convergence in *r*th Mean Implies Convergence in Probability

Theorem. *If* $X_n \xrightarrow{rth} X$, *then* $X_n \xrightarrow{p} X$.

PROOF. Using the indicator function notation of **1.1.10** we have, for any $\varepsilon > 0$,

$$E|X_n - X|^r \geq E\{|X_n - X|^r I(|X_n - X| > \varepsilon)\} \geq \varepsilon^r P(|X_n - X| > \varepsilon)$$

and thus

$$P(|X_n - X| > \varepsilon) \leq \varepsilon^{-r} E|X_n - X|^r \to 0, n \to \infty. \quad \blacksquare$$

1.3.3 Convergence in Probability Implies Convergence in Distribution

(This will be proved in Section 1.5, but is stated here for completeness.)

1.3.4 Convergence in Probability Sufficiently Fast Implies Convergence *wp*1

Theorem. *If*

$$(*) \qquad \sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) < \infty \qquad \textit{for every} \quad \varepsilon > 0,$$

then $X_n \xrightarrow{wp1} X$.

PROOF. Let $\varepsilon > 0$ be given. We have

$$(**) \qquad P(|X_m - X| > \varepsilon \text{ for some } m \geq n) = P\left(\bigcup_{m=n}^{\infty} \{|X_m - X| > \varepsilon\}\right)$$

$$\leq \sum_{m=n}^{\infty} P(|X_m - X| > \varepsilon).$$

Since the sum in (**) is the tail of a convergent series and hence $\to 0$ as $n \to \infty$, the alternate condition for convergence *wp*1 follows. $\blacksquare$

Note that the condition of the theorem defines a mode of convergence stronger than convergence *wp*1. Following Hsu and Robbins (1947), we say that X_n *converges completely* to X if (*) holds.

1.3.5 Convergence in rth Mean Sufficiently Fast Implies Convergence *wp*1

The preceding result, in conjunction with the proof of Theorem 1.3.2, yields

Theorem. *If* $\sum_{n=1}^{\infty} \mathrm{E}|\mathrm{X}_n - \mathrm{X}|^r < \infty$, *then* $\mathrm{X}_n \xrightarrow{\text{wp1}} \mathrm{X}$.

The hypothesis of the theorem in fact yields the much stronger conclusion that the random series $\sum_{n=1}^{\infty} |X_n - X|^r$ converges *wp*1 (see Lukacs (1975), Section 4.2, for details).

1.3.6 Dominated Convergence in Probability Implies Convergence in Mean

Theorem. *Suppose that* $\mathrm{X}_n \xrightarrow{p} \mathrm{X}$, $|\mathrm{X}_n| \le |\mathrm{Y}|$ wp1 (*all* n), *and* $\mathrm{E}|\mathrm{Y}|^r < \infty$. *Then* $\mathrm{X}_n \xrightarrow{r\text{th}} \mathrm{X}$.

PROOF. First let us check that $|X| \le |Y|$ *wp*1. Given $\delta > 0$, we have $P(|X| > |Y| + \delta) \le P(|X| > |X_n| + \delta) \le P(|X_n - X| > \delta) \to 0$, $n \to \infty$. Hence $|X| \le |Y| + \delta$ *wp*1 for any $\delta > 0$ and so for $\delta = 0$.

Consequently, $|X_n - X| \le |X| + |X_n| \le 2|Y|$ *wp*1.

Now choose and fix $\varepsilon > 0$. Since $E|Y|^r < \infty$, there exists a finite constant $A_\varepsilon > \varepsilon$ such that $E\{|Y|^r I(2|Y| > A_\varepsilon)\} \le \varepsilon$. We thus have

$$\begin{aligned} E|X_n - X|^r &= E\{|X_n - X|^r I(|X_n - X| > A_\varepsilon)\} \\ &\quad + E\{|X_n - X|^r I(|X_n - X| \le \varepsilon)\} \\ &\quad + E\{|X_n - X|^r I(\varepsilon < |X_n - X| \le A_\varepsilon)\} \\ &\le E\{(|2Y|^r I(2|Y| > A_\varepsilon)\} + \varepsilon^r + A_\varepsilon^r P(|X_n - X| > \varepsilon) \\ &\le 2^r\varepsilon + \varepsilon^r + A_\varepsilon^r P(|X_n - X| > \varepsilon). \end{aligned}$$

Since $P(|X_n - X| > \varepsilon) \to 0$, $n \to \infty$, the right-hand side becomes less than $2^r\varepsilon + 2\varepsilon^r$ for all n sufficiently large. ■

More general theorems of this type are discussed in Section 1.4.

1.3.7 Dominated Convergence *wp*1 Implies Convergence in Mean

By 1.3.1 we may replace $\xrightarrow{p}$ by $\xrightarrow{wp1}$ in Theorem 1.3.6, obtaining

Theorem. *Suppose that* $\mathrm{X}_n \xrightarrow{wp1} \mathrm{X}$, $|\mathrm{X}_n| \le |\mathrm{Y}|$ wp1 (*all* n), *and* $\mathrm{E}|\mathrm{Y}|^r < \infty$. *Then* $\mathrm{X}_n \xrightarrow{r\text{th}} \mathrm{X}$.

1.3.8 Some Counterexamples

Sequences $\{X_n\}$ convergent in probability but *not* *wp*1 are provided in Examples A, B and C. The sequence in Example B is also convergent in mean square. A sequence convergent in probability but *not* in rth mean for any $r > 0$ is provided in Example D. Finally, to obtain a sequence convergent

*wp*1 but *not* in *r*th mean for any $r > 0$, take an appropriate subsequence of the sequence in Example D (Problem 1.P.6). For more counterexamples, see Chung (1974), Section 4.1, and Lukacs (1975), Section 2.2, and see Section 2.1.

Example A. The usual textbook examples are versions of the following (Royden (1968), p. 92). Let $(\Omega, \mathscr{A}, P)$ be the probability space corresponding to Ω the interval $[0, 1]$, $\mathscr{A}$ the Borel sets in $[0, 1]$, and P the Lebesgue measure on $\mathscr{A}$. For each $n = 1, 2, \ldots$, let k_n and v_n satisfy $n = k_n + 2^{v_n}$, $0 \le k_n < 2^{v_n}$, and define

$$X_n(\omega) = \begin{cases} 1, & \text{if } \omega \in [k_n 2^{-v_n}, (k_n + 1)2^{-v_n}] \\ 0, & \text{otherwise.} \end{cases}$$

It is easily seen that $X_n \xrightarrow{p} 0$ yet $X_n(\omega) \to 0$ holds *nowhere*, $\omega \in [0, 1]$. ■

Example B. Let $Y_1, Y_2, \ldots$ be I.I.D. random variables with mean 0 and variance 1. Define

$$X_n = \frac{\sum_1^n Y_i}{(n \log \log n)^{1/2}}.$$

By the central limit theorem (Section 1.9) and theorems presented in Section 1.5, it is clear that $X_n \xrightarrow{p} 0$. Also, by direct computation, it is immediate that $X_n \xrightarrow{2nd} 0$. However, by the law of the iterated logarithm (Section 1.10), it is evident that $X_n(\omega) \to 0$, $n \to \infty$, only for ω in a set of probability 0. ■

Example C (contributed by J. Sethuraman). Let $Y_1, Y_2, \ldots$ be I.I.D. random variables. Define $X_n = Y_n/n$. Then clearly $X_n \xrightarrow{p} 0$. However, $X_n \xrightarrow{wp1} 0$ if and only if $E|Y_1| < \infty$. To verify this claim, we apply

Lemma (Chung (1974), Theorem 3.2.1). *For any positive random variable* Z,

$$\sum_{n=1}^{\infty} P(Z \ge n) \le E\{Z\} \le 1 + \sum_{n=1}^{\infty} P(Z \ge n).$$

Thus, utilizing the identical distributions assumption, we have

$$\sum_{n=1}^{\infty} P(|X_n| \ge \varepsilon) = \sum_{n=1}^{\infty} P(|Y_1| \ge n\varepsilon) \le \frac{1}{\varepsilon} E|Y_1|,$$

$$1 + \sum_{n=1}^{\infty} P(|X_n| \ge \varepsilon) = 1 + \sum_{n=1}^{\infty} P(|Y_1| \ge n\varepsilon) \ge \frac{1}{\varepsilon} E|Y_1|.$$

The result now follows, with the use of the independence assumption, by an application of the Borel–Cantelli lemma (Appendix). ■

Example D. Consider

$$X_n = \begin{cases} n, & \text{with probability } 1/\log n \\ 0, & \text{with probability } 1\text{-}1/\log n. \end{cases}$$

Clearly $X_n \xrightarrow{p} 0$. However, for any $r > 0$,

$$E|X_n|^r = \frac{n^r}{\log n} \to \infty. \quad \blacksquare$$

1.4 CONVERGENCE OF MOMENTS; UNIFORM INTEGRABILITY

Suppose that X_n converges to X in one of the senses $\xrightarrow{d}$, $\xrightarrow{p}$, $\xrightarrow{wp1}$ or $\xrightarrow{rth}$. What is implied regarding convergence of $E\{X_n^s\}$ to $E\{X^s\}$, or $E|X_n|^s$ to $E|X|^s, n \to \infty$? The basic answer is provided by Theorem A, in the general context of $\xrightarrow{d}$, which includes the other modes of convergence. Also, however, specialized results are provided for the cases $\xrightarrow{rth}$, $\xrightarrow{p}$, and $\xrightarrow{wp1}$. These are given by Theorems B, C, and D, respectively.

Before proceeding to these results, we introduce three special notions and examine their interrelationships. A sequence of random variables $\{Y_n\}$ is *uniformly integrable* if

$$\lim_{c \to \infty} \sup_n E\{|Y_n| I(|Y_n| > c)\} = 0.$$

A sequence of set functions $\{Q_n\}$ defined on $\mathcal{A}$ is *uniformly absolutely continuous* with respect to a measure P on $\mathcal{A}$ if, given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$P(A) < \delta \Rightarrow \sup_n |Q_n(A)| < \varepsilon.$$

The sequence $\{Q_n\}$ is *equicontinuous at* ϕ if, given $\varepsilon > 0$ and a sequence $\{A_n\}$ in $\mathcal{A}$ decreasing to ϕ, there exists M such that

$$m > M \Rightarrow \sup_n |Q_n(A_m)| < \varepsilon.$$

Lemma A. (i) *Uniform integrability of* $\{Y_n\}$ *on* $(\Omega, \mathcal{A}, P)$ *is equivalent to the pair of conditions*

(a) $\sup_n E|Y_n| < \infty$
and

(b) *the set functions* $\{Q_n\}$ *defined by* $Q_n(A) = \int_A |Y_n| dP$ *are uniformly absolutely continuous with respect to* P.

(ii) *Sufficient for uniform integrability of* $\{Y_n\}$ *is that*

$$\sup_n E|Y_n|^{1+\varepsilon} < \infty$$

for some $\varepsilon > 0$.

(iii) *Sufficient for uniform integrability of* $\{Y_n\}$ *is that there be a random variable* Y *such that* $E|Y| < \infty$ *and*

$$P(|Y_n| \geq y) \leq P(|Y| \geq y), \text{ all } n \geq 1, \text{ all } y > 0.$$

(iv) *For set functions* Q_n *each absolutely continuous with respect to a measure* P, *equicontinuity at* ϕ *implies uniform absolute continuity with respect to* P.

PROOF. (i) Chung (1974), p. 96; (ii) note that

$$E\{|Y_n| I(|Y_n| > c)\} \leq c^{-\varepsilon} E|Y_n|^{1+\varepsilon};$$

(iii) Billingsley (1968), p. 32; (iv) Kingman and Taylor (1966), p. 178. ■

Theorem A. *Suppose that* $X_n \xrightarrow{d} X$ *and the sequence* $\{X_n^r\}$ *is uniformly integrable, where* $r > 0$. *Then* $E|X|^r < \infty$, $\lim_n E\{X_n^r\} = E\{X^r\}$, *and* $\lim_n E|X_n|^r = E|X|^r$.

PROOF. Denote the distribution function of X by F. Let $\varepsilon > 0$ be given. Choose c such that $\pm c$ are continuity points of F and, by the uniform integrability, such that

$$\sup_n E\{|X_n|^r I(|X_n| \geq c)\} < \varepsilon.$$

For any $d > c$ such that $\pm d$ are also continuity points of F, we obtain from the second theorem of Helly (Appendix) that

$$\lim_{n\to\infty} E\{|X_n|^r I(c \leq |X_n| \leq d)\} = E\{|X|^r I(c \leq |X| \leq d)\}.$$

It follows that $E\{|X|^r I(c \leq |X| \leq d)\} < \varepsilon$ for all such choices of d. Letting $d \to \infty$, we obtain $E\{|X|^r I(|X| \geq c)\} < \varepsilon$, whence $E|X|^r < \infty$.

Now, for the same c as above, write

$$\begin{aligned} |E\{X_n^r\} - E\{X^r\}| \leq\ & |E\{X_n^r I(|X_n| \leq c)\} - E\{X^r I(|X| \leq c)\}| \\ & + E\{|X_n|^r I(|X_n| > c)\} + E\{|X|^r I(|X| > c)\}. \end{aligned}$$

By the Helly theorem again, the first term on the right-hand side tends to 0 as $n \to \infty$. The other two terms on the right are each less than ε. Thus $\lim_n E\{X_n^r\} = E\{X^r\}$. A similar argument yields $\lim_n E|X_n|^r = E|X|^r$. ■

By arguments similar to the preceding, the following partial converse to Theorem A may be obtained (Problem 1.P.7).

Lemma B. Suppose that $X_n \xrightarrow{d} X$ *and* $\lim_n E|X_n|^r = E|X|^r < \infty$. *Then the sequence* $\{X_n^r\}$ *is uniformly integrable.*

We now can easily establish a simple theorem apropos to the case $\xrightarrow{r\text{th}}$.

Theorem B. Suppose that $X_n \xrightarrow{r\text{th}} X$ *and* $E|X|^r < \infty$. *Then* $\lim_n E\{X_n^r\} = E\{X^r\}$ *and* $\lim_n E|X_n|^r = E|X|^r$.

PROOF. For $0 < r \leq 1$, apply the inequality $|x + y|^r \leq |x|^r + |y|^r$ to write $\|x|^r - |y|^r| \leq |x - y|^r$ and thus

$$|E|X_n|^r - E|X|^r| \leq E|X_n - X|^r.$$

For $r > 1$, apply Minkowski's inequality (Appendix) to obtain

$$|(E|X_n|^r)^{1/r} - (E|X|^r)^{1/r}| \leq (E|X_n - X|^r)^{1/r}.$$

In either case, $\lim_n E|X_n|^r = E|X|^r < \infty$ follows. Therefore, by Lemma B, $\{X_n^r\}$ is uniformly integrable. Hence, by Theorem A, $\lim_n E\{X_n^r\} = E\{X^r\}$ follows. ■

Next we present results oriented to the case $\xrightarrow{p}$.

Lemma C. Suppose that $X_n \xrightarrow{p} X$ *and* $E|X_n|^r < \infty$, *all* n. *Then the following statements hold.*

(i) $X_n \xrightarrow{r\text{th}} X$ *if and only if the sequence* $\{X_n^r\}$ *is uniformly integrable.*

(ii) *If the set functions* $\{Q_n\}$ *defined by* $Q_n(A) = \int_A |X_n|^r \, dP$ *are equicontinuous at* ϕ, *then* $X_n \xrightarrow{r\text{th}} X$ *and* $E|X|^r < \infty$.

PROOF. (i) see Chung (1974), pp. 96–97; (ii) see Kingman and Taylor (1966), pp. 178–180. ■

It is easily checked (Problem 1.P.8) that each of parts (i) and (ii) generalizes Theorem 1.3.6.

Combining Lemma C with Theorem B and Lemma A, we have

Theorem C. Suppose that $X_n \xrightarrow{p} X$ *and that either*

(i) $E|X|^r < \infty$ *and* $\{X_n^r\}$ *is uniformly integrable,*

or

(ii) $\sup_n E|X_n|^r < \infty$ *and the set functions* $\{Q_n\}$ *defined by* $Q_n(A) = \int_A |X_n|^r \, dP$ *are equicontinuous at* ϕ.

Then $\lim_n E\{X_n^r\} = E\{X^r\}$ *and* $\lim_n E|X_n|^r = E|X|^r$.

Finally, for the case $\xrightarrow{wp1}$, the preceding result may be used; but also, by a simple application (Problem 1.P.9) of Fatou's lemma (Appendix), the following is easily obtained.

Theorem D. Suppose that $X_n \xrightarrow{wp1} X$. *If* $\overline{\lim}_n E|X_n|^r \leq E|X|^r < \infty$, *then* $\lim_n E\{X_n^r\} = E\{X^r\}$ *and* $\lim_n E|X_n|^r = E|X|^r$.

As noted at the outset of this section, the fundamental result on convergence of moments is provided by Theorem A, which imposes a uniform integrability condition. For practical implementation of the theorem, Lemma A (i), (ii), (iii) provides various sufficient conditions for uniform integrability. Justification for the trouble of verifying uniform integrability is provided by Lemma B, which shows that the uniform integrability condition is essentially necessary.

1.5 FURTHER DISCUSSION OF CONVERGENCE IN DISTRIBUTION

This mode of convergence has been treated briefly in Sections 1.2–1.4. Here we provide a collection of basic facts about it. Recall that the definition of $X_n \xrightarrow{d} X$ is expressed in terms of the corresponding distribution functions F_n and F, and that the alternate notation $F_n \Rightarrow F$ is often convenient. The reader should formulate "convergence in distribution" for random *vectors.*

1.5.1 Criteria for Convergence in Distribution

The following three theorems provide *methodology* for establishing convergence in distribution.

Theorem A. Let the distribution functions $F, F_1, F_2, \ldots$ *possess respective characteristic functions* $\phi, \phi_1, \phi_2, \ldots$. *The following statements are equivalent*:

(i) $F_n \Rightarrow F$;

(ii) $\lim_n \phi_n(t) = \phi(t)$, *each real* t;

(iii) $\lim_n \int g\, dF_n = \int g\, dF$, *each bounded continuous function* g.

PROOF. That (i) implies (iii) is given by the generalized Helly theorem (Appendix). We now show the converse. Let t be a continuity point of F and let $\varepsilon > 0$ be given. Take any continuous function g satisfying $g(x) = 1$ for $x \leq t$, $0 \leq g(x) \leq 1$ for $t < x < t + \varepsilon$, and $g(x) = 0$ for $x \geq t + \varepsilon$. Then, assuming (iii), we obtain (Problem 1.P.10)

$$\overline{\lim_{n \to \infty}} F_n(t) \leq F(t + \varepsilon).$$

Similarly, (iii) also gives

$$\underline{\lim}_{n\to\infty} F_n(t) \geq F(t - \varepsilon).$$

Thus (i) follows.

For proof that (i) and (ii) are equivalent, see Gnedenko (1962), p. 285. ■

Example. If the characteristic function of a random variable X_n tends to the function $\exp(-\frac{1}{2}t^2)$ as $n \to \infty$, then $X_n \xrightarrow{d} N(0, 1)$. ■

The multivariate version of Theorem A is easily formulated.

Theorem B (Fréchet and Shohat). *Let the distribution functions* F_n *possess finite moments* $\alpha_k^{(n)} = \int t^k \, dF_n(t)$ *for* $k = 1, 2, \ldots$ *and* $n = 1, 2, \ldots$. *Assume that the limits* $\alpha_k = \lim_n \alpha_k^{(n)}$ *exist (finite), each* k. *Then*

(i) *the limits* $\{\alpha_k\}$ *are the moments of a distribution function* F;

(ii) *if the* F *given by* (i) *is unique, then* $F_n \Rightarrow F$.

For proof, see Fréchet and Shohat (1931), or Loève (1977), Section 11.4. This result provides a convergence of moments criterion for convergence in distribution. In implementing the criterion, one would also utilize Theorem 1.13, which provides conditions under which the moments $\{\alpha_k\}$ determine a unique F.

The following result, due to Scheffé (1947), provides a convergence of densities criterion. (See Problem 1.P.11.)

Theorem C (Scheffé). *Let* $\{f_n\}$ *be a sequence of densities of absolutely continuous distribution functions, with* $\lim_n f_n(x) = f(x)$, *each real* x. *If* f *is a density function, then* $\lim_n \int |f_n(x) - f(x)| dx = 0$.

PROOF. Put $g_n(x) = [f(x) - f_n(x)]I(f(x) \geq f_n(x))$, each x. Using the fact that f is a density, check that

$$\int |f_n(x) - f(x)| dx = 2 \int g_n(x) dx.$$

Now $|g_n(x)| \leq f(x)$, all x, each n. Hence, by dominated convergence (Theorem 1.3.7), $\lim_n \int g_n(x) dx = 0$. ■

1.5.2 Reduction of Multivariate Case to Univariate Case

The following result, due to Cramér and Wold (1936), allows the question of convergence of multivariate distribution functions to be reduced to that of convergence of univariate distribution functions.

Theorem. *In $\mathbf{R}^k$, the random vectors $\mathbf{X}_n$ converge in distribution to the random vector $\mathbf{X}$ if and only if each linear combination of the components of $\mathbf{X}_n$ converges in distribution to the same linear combination of the components of $\mathbf{X}$.*

PROOF. Put $\mathbf{X}_n = (X_{n1}, \ldots, X_{nk})$ and $\mathbf{X} = (X_1, \ldots, X_k)$ and denote the corresponding characteristic functions by ϕ_n and ϕ. Assume now that for any real $\lambda_1, \ldots, \lambda_k$,

$$\lambda_1 X_{n1} + \cdots + \lambda_k X_{nk} \xrightarrow{d} \lambda_1 X_1 + \cdots + \lambda_k X_k.$$

Then, by Theorem 1.5.1A,

$$\lim_{n\to\infty} \phi_n(t\lambda_1, \ldots, t\lambda_k) = \phi(t\lambda_1, \ldots, t\lambda_k), \text{ all } t.$$

With $t = 1$, and since $\lambda_1, \ldots, \lambda_k$ are arbitrary, it follows by the multivariate version of Theorem 1.5.1A that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.

The converse is proved by a similar argument. ■

Some extensions due to Wald and Wolfowitz (1944) and to Varadarajan (1958) are given in Rao (1973), p. 128. Also, see Billingsley (1968), p. 49, for discussion of this "Cramer–Wold device."

1.5.3 Uniformity of Convergence in Distribution

An important question regarding the weak convergence of F_n to F is whether the pointwise convergences hold uniformly. The following result is quite useful.

Theorem (Pólya). *If $\mathbf{F}_n \Rightarrow \mathbf{F}$ and $\mathbf{F}$ is continuous, then*

$$\lim_{n\to\infty} \sup_t |\mathbf{F}_n(t) - \mathbf{F}(t)| = 0.$$

The proof is left as an exercise (Problem 1.P.12). For generalities, see Ranga Rao (1962).

1.5.4 Convergence in Distribution for Perturbed Random Variables

A common situation in mathematical statistics is that the statistic of interest is a slight modification of a random variable having a known limit distribution. A fundamental role is played by the following theorem, which was developed by Slutsky (1925) and popularized by Cramér (1946). Note that no restrictions are imposed on the possible dependence among the random variables involved.

Theorem (Slutsky). *Let* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$, *where* c *is a finite constant. Then*

(i) $X_n + Y_n \xrightarrow{d} X + c$;

(ii) $X_n Y_n \xrightarrow{d} cX$;

(iii) $X_n/Y_n \xrightarrow{d} X/c$ *if* $c \neq 0$.

Corollary A. *Convergence in probability,* $X_n \xrightarrow{p} X$, *implies convergence in distribution,* $X_n \xrightarrow{d} X$.

Corollary B. *Convergence in probability to a constant is equivalent to convergence in distribution to the given constant.*

Note that Corollary A was given previously in **1.3.3**. The method of proof of the theorem is demonstrated sufficiently by proving (i). The proofs of (ii) and (iii) and of the corollaries are left as exercises (see Problems 1.P.13–14).

PROOF OF (i). Choose and fix t such that $t - c$ is a continuity point of F_X. Let $\varepsilon > 0$ be such that $t - c + \varepsilon$ and $t - c - \varepsilon$ are also continuity points of F_X. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \leq t) \\ &\leq P(X_n + Y_n \leq t, |Y_n - c| < \varepsilon) + P(|Y_n - c| \geq \varepsilon) \\ &\leq P(X_n \leq t - c + \varepsilon) + P(|Y_n - c| \geq \varepsilon). \end{aligned}$$

Hence, by the hypotheses of the theorem, and by the choice of $t - c + \varepsilon$,

$$(*) \qquad \overline{\lim_n} F_{X_n+Y_n}(t) \leq \overline{\lim_n} P(X_n \leq t - c + \varepsilon) + \overline{\lim_n} P(|Y_n - c| \geq \varepsilon)$$

$$= F_X(t - c + \varepsilon).$$

Similarly,

$$P(X_n \leq t - c - \varepsilon) \leq P(X_n + Y_n \leq t) + P(|Y_n - c| \geq \varepsilon)$$

and thus

$$(**) \qquad F_X(t - c - \varepsilon) \leq \underline{\lim_n} F_{X_n+Y_n}(t).$$

Since $t - c$ is a continuity point of F_X, and since ε may be taken arbitrarily small, (*) and (**) yield

$$\lim_n F_{X_n+Y_n}(t) = F_X(t - c) = F_{X+c}(t). \quad \blacksquare$$

1.5.5 Asymptotic Normality

The most important special case of convergence in distribution consists of convergence to a normal distribution. A sequence of random variables $\{X_n\}$ converges in distribution to $N(\mu, \sigma^2)$, $\sigma > 0$, if equivalently, the sequence $\{(X_n - \mu)/\sigma\}$ converges in distribution to $N(0, 1)$. (Verify by Slutsky's Theorem.)

More generally, a sequence of random variables $\{X_n\}$ is *asymptotically normal* with "mean" μ_n and "variance" σ_n^2 if $\sigma_n > 0$ for all n sufficiently large and

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0, 1).$$

We write "X_n is $AN(\mu_n, \sigma_n^2)$." Here $\{\mu_n\}$ and $\{\sigma_n\}$ are sequences of constants. It is not necessary that μ_n and σ_n^2 be the mean and variance of X_n, nor even that X_n possess such moments. Note that if X_n is $AN(\mu_n, \sigma_n^2)$, it does not necessarily follow that $\{X_n\}$ converges in distribution to anything. Nevertheless in any case we have (show why)

$$\sup_t |P(X_n \le t) - P(N(\mu_n, \sigma_n^2) \le t)| \to 0, \qquad n \to \infty,$$

so that for a range of probability calculations we may treat X_n as a $N(\mu_n, \sigma_n^2)$ random variable.

As exercises (Problems 1.P.15–16), prove the following useful lemmas.

Lemma A. *If* X_n *is* $AN(\mu_n, \sigma_n^2)$, *then also* X_n *is* $AN(\bar{\mu}_n, \bar{\sigma}_n^2)$ *if and only if*

$$\frac{\bar{\sigma}_n}{\sigma_n} \to 1, \frac{\bar{\mu}_n - \mu_n}{\sigma_n} \to 0.$$

Lemma B. *If* X_n *is* $AN(\mu_n, \sigma_n^2)$, *then also* $a_n X_n + b_n$ *is* $AN(\mu_n, \sigma_n^2)$ *if and only if*

$$a_n \to 1, \frac{\mu_n(a_n - 1) + b_n}{\sigma_n} \to 0.$$

Example. If X_n is $AN(n, 2n)$, then so is

$$\frac{n-1}{n} X_n$$

but not

$$\frac{\sqrt{n} - 1}{\sqrt{n}} X_n. \quad \blacksquare$$

We say that a sequence of random *vectors* $\{\mathbf{X}_n\}$ is *asymptotically (multivariate) normal* with "mean vector" $\boldsymbol{\mu}_n$ and "covariance matrix" $\boldsymbol{\Sigma}_n$ if $\boldsymbol{\Sigma}_n$ has nonzero diagonal elements for all n sufficiently large, and for every vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}\boldsymbol{\Sigma}_n\boldsymbol{\lambda}' > 0$ for all n sufficiently large, the sequence $\boldsymbol{\lambda}\mathbf{X}_n'$ is $AN(\boldsymbol{\lambda}\boldsymbol{\mu}_n', \boldsymbol{\lambda}\boldsymbol{\Sigma}_n\boldsymbol{\lambda}')$. We write "$\mathbf{X}_n$ is $AN(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$." Here $\{\boldsymbol{\mu}_n\}$ is a sequence of vector constants and $\{\boldsymbol{\Sigma}_n\}$ a sequence of covariance matrix constants. As an exercise (Problem 1.P.17), show that $\mathbf{X}_n$ is $AN(\boldsymbol{\mu}_n, c_n^2\boldsymbol{\Sigma})$ if and only if

$$\frac{\mathbf{X}_n - \boldsymbol{\mu}_n}{c_n} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Here $\{c_n\}$ is a sequence of real constants and $\boldsymbol{\Sigma}$ a covariance matrix.

1.5.6 Inverse Functions of Weakly Convergent Distributions

The following result will be utilized in Section 1.6 in proving Theorem 1.6.3.

Lemma. *If* $\mathrm{F}_n \Rightarrow \mathrm{F}$, *then the set*

$$\{t: 0 < t < 1, \mathrm{F}_n^{-1}(t) \not\to \mathrm{F}^{-1}(t), n \to \infty\}$$

contains at most countably many elements.

PROOF. Let $0 < t_0 < 1$ be such that $F_n^{-1}(t_0) \not\to F^{-1}(t_0), n \to \infty$. Then there exists an $\varepsilon > 0$ such that $F^{-1}(t_0) \pm \varepsilon$ are continuity points of F and $|F_n^{-1}(t_0) - F^{-1}(t_0)| > \varepsilon$ for infinitely many $n = 1, 2, \ldots$. Suppose that $F_n^{-1}(t_0) < F^{-1}(t_0) - \varepsilon$ for infinitely many n. Then, by Lemma 1.1.4(ii), $t_0 \le F_n(F_n^{-1}(t_0)) \le F_n(F^{-1}(t_0) - \varepsilon)$. Thus the convergence $F_n \Rightarrow F$ yields $t_0 \le F(F^{-1}(t_0) - \varepsilon)$, which in turn yields, by Lemma 1.1.4(i), $F^{-1}(t_0) \le F^{-1}(F(F^{-1}(t_0) - \varepsilon)) \le F^{-1}(t_0) - \varepsilon$, a contradiction. Therefore, we must have

$$F_n^{-1}(t_0) > F^{-1}(t_0) + \varepsilon \qquad \text{for infinitely many} \quad n = 1, 2, \ldots.$$

By Lemma 1.1.4(iii), this is equivalent to

$$F_n(F^{-1}(t_0) + \varepsilon) < t_0 \qquad \text{for infinitely many} \quad n = 1, 2, \ldots,$$

which by the convergence $F_n \Rightarrow F$ yields $F(F^{-1}(t_0) + \varepsilon) \le t_0$. But also $t_0 \le F(F^{-1}(t_0))$, by Lemma 1.1.4(i). It follows that

$$t_0 = F(F^{-1}(t_0))$$

and that

$$F(x) = t_0 \qquad \text{for} \quad x \in [F^{-1}(t_0), F^{-1}(t_0) + \varepsilon],$$