A COURSE IN LARGE SAMPLE THEORY

Thomas S. Ferguson

Texts in Statistical Science

Springer-Science+Business Media, B.V.

A Course in Large Sample Theory

TEXTS IN STATISTICAL SCIENCE SERIES

OTHER TITLES IN THE SERIES INCLUDE:

Computer-Aided Multivariate Analysis
Third edition
A. Afifi and V. Clark

Practical Statistics for Medical Research
D.G. Altman

Interpreting Data
A. J. B. Anderson

Statistical Methods for SPC and TQM
D. Bissell

Statistics in Research and Development
Second edition
R. Caulcutt

The Analysis of Time Series – An introduction
Fifth edition
C. Chatfield

Problem Solving – A statistician's guide
Second edition
C. Chatfield

Introduction to Multivariate Analysis
C. Chatfield and A. J. Collins

Modelling Binary Data
D. Collett

Modelling Survival Data in Medical Research
D. Collett

Applied Statistics
D. R. Cox and E. J. Snell

Statistical Analysis of Reliability Data
M. J. Crowder, A. C. Kimber, T. J. Sweeting and R. L. Smith

An Introduction to Generalized Linear Models
A. J. Dobson

Introduction to Optimization Methods and their Applications in Statistics
B. S. Everitt

Multivariate Statistics – A practical approach
B. Flury and H. Riedwyl

Readings in Decision Analysis
S. French

Practical Longitudinal Data Analysis
D. Hand and M. Crowder

Multivariate Analysis of Variance and Repeated Measures
D. J. Hand and C. C. Taylor

The Theory of Linear Models
B. Jørgensen

Modeling and Analysis of Stochastic Systems
V. G. Kulkarni

Statistics for Accountants
S. Letchford

Statistical Theory
Fourth edition
B. Lindgren

Randomization and Monte Carlo Methods in Biology
B. F. J. Manly

Statistical Methods in Agriculture and Experimental Biology
Second edition
R. Mead, R. N. Curnow and A. M. Hasted

Statistics in Engineering
A. V. Metcalfe

Elements of Simulation
B. J. T. Morgan

Probability – Methods and measurement
A. O'Hagan

Essential Statistics
Second edition
D. G. Rees

Large Sample Methods in Statistics
P. K. Sen and J. M. Singer

Decision analysis – A Bayesian Approach
J. Q. Smith

Applied Nonparametric Statistical Methods
Second edition
P. Sprent

Elementary Applications of Probability Theory
Second edition
H. C. Tuckwell

Statistical Process Control – Theory and practice
Third edition
G. B. Wetherill and D. W. Brown

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Full information on the complete range of Chapman & Hall statistics books is available from the publishers.

A Course in Large Sample Theory

Thomas S. Ferguson
Professor of Statistics
University of California
Los Angeles
USA

Springer-Science+Business Media, B.V.

First edition 1996

Originally published by Chapman & Hall 1996

Typeset in the USA by Brookhaven Typesetting Systems, Brookhaven, New York

ISBN 978-0-412-04371-0 ISBN 978-1-4899-4549-5 (eBook)
DOI 10.1007/978-1-4899-4549-5

A catalogue record for this book is available from the British Library

∞ Printed on permanent acid-free text paper, manufactured in accordance with ANSI/NISO Z39.48-1992 and ANSI/NISO Z39.48-1984 (Permanence of Paper).

Contents

Preface

The subject area of mathematical statistics is so vast that in undergraduate courses there is only time enough to present an overview of the material. In particular, proofs of theorems are often omitted, occasionally with a reference to specialized material, with the understanding that proofs will be given in later, presumably graduate, courses. Some undergraduate texts contain an outline of the proof of the central limit theorem, but other theorems useful in the large sample analysis of statistical problems are usually stated and used without proof. Typical examples concern topics such as the asymptotic normality of the maximum likelihood estimate, the asymptotic distribution of Pearson's chi-square statistic, the asymptotic distribution of the likelihood ratio test, and the asymptotic normality of the rank-sum test statistic.

But then in graduate courses, it often happens that proofs of theorems are assumed to be given in earlier, possibly undergraduate, courses, or proofs are given as they arise in specialized settings. Thus the student never learns in a general methodical way one of the most useful areas for research in statistics – large sample theory, or as it is also called, asymptotic theory. There is a need for a separate course in large sample theory at the beginning graduate level. It is hoped that this book will help in filling this need.

A course in large sample theory has been given at UCLA as the second quarter of our basic graduate course in theoretical statistics for about twenty years. The students who have learned large sample theory by the route given in this text can be said to form a large sample. Although this course is given in the Mathematics Department, the clients have been a mix of graduate students from various disciplines. Roughly 40% of the students have been from Mathematics, possibly 30% from Biostatistics, and the rest from Biomathematics, Engineering, Economics, Business, and other fields. The students generally find the course challenging and interesting, and have often contributed to the improvement of the course through questions, suggestions and, of course, complaints.

Because of the mix of students, the mathematical background required for the course has necessarily been restricted. In particular, it could not be assumed that the students have a background in measure-theoretic analysis or probability. However, for an understanding of this book, an undergraduate course in analysis is needed as well as a good undergraduate course in mathematical statistics.

Statistics is a multivariate discipline. Nearly, every useful univariate problem has important multivariate extensions and applications. For this reason, nearly all theorems are stated in a multivariate setting. Often the statement of a multivariate theorem is identical to the univariate version, but when it is not, the reader may find it useful to consider the theorem carefully in one dimension first, and then look at the examples and exercises that treat problems in higher dimensions.

The material is constructed in consideration of the student who wants to learn techniques of large sample theory on his/her own without the benefit of a classroom environment. There are many exercises, and solutions to all exercises may be found in the appendix. For use by instructors, other exercises, without solutions, can be found on the web page for the course, at http://www.stat.ucla.edu/courses/graduate/M276B/.

Each section treats a specific topic and the basic idea or central result of the section is stated as a theorem. There are 24 sections and so there are 24 theorems. The sections are grouped into four parts. In the first part, basic notions of limits in probability theory are treated, including laws of large numbers and the central limit theorem. In the second part, certain basic tools in statistical asymptotic theory, such as Slutsky's Theorem and Cramér's Theorem, are discussed and illustrated, and finally used to derive the asymptotic distribution and power of Pearson's chi-square. In the third part, certain special topics are treated by the methods of the first two parts, such as some time series statistics, some rank statistics, and distributions of quantiles and extreme order statistics. The last part contains a treatment of standard statistical techniques including maximum likelihood estimation, the likelihood ratio test, asymptotic normality of Bayes estimates, and minimum chi-square estimation. Parts 3 and 4 may be read independently. There is easily enough material in the book for a semester course. In a quarter course, some material in parts 3 and 4 will have to be omitted or skimmed.

I would like to acknowledge a great debt this book owes to Lucien Le Cam not only for specific details as one may note in references to him in the text here and there, but also for a general philosophic outlook on the

subject. Since the time I learned the subject from him many years ago, he has developed a much more general and mathematical approach to the subject that may be found in his book, Le Cam (1986) mentioned in the references.

Rudimentary versions of this book in the form of notes have been in existence for some 20 years, and have undergone several changes in computer systems and word processors. I am indebted to my wife, Beatriz, for cheerfully typing some of these conversions. Finally, I am indebted to my students, too numerous to mention individually. Each class was distinctive and each class taught me something new so that the next year's class was taught somewhat differently than the last. If future students find this book helpful, they also can thank these students for their contribution to making it understandable.

Thomas S. Ferguson, April 1996

1

Basic Probability Theory

1

Modes of Convergence

We begin by studying the relationships among four distinct modes of convergence of a sequence of random vectors to a limit. All convergences are defined for d-dimensional random vectors. For a random vector $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$, the distribution function of $\mathbf{X}$, defined for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, is denoted by $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$. The Euclidean norm of $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is denoted by $|\mathbf{x}| = (x_1^2 + \cdots + x_d^2)^{1/2}$. Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ be random vectors with values in $\mathbb{R}^d$.

Definition 1. $\mathbf{X}_n$ converges *in law* to $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{\mathscr{L}} \mathbf{X}$, if $F_{\mathbf{X}_n}(\mathbf{x}) \longrightarrow F_{\mathbf{X}}(\mathbf{x})$ as $n \to \infty$, for all points $\mathbf{x}$ at which $F_{\mathbf{X}}(\mathbf{x})$ is continuous.

Convergence in law is the mode of convergence most used in the following chapters. It is the mode found in the Central Limit Theorem and is sometimes called convergence *in distribution*, or *weak* convergence.

Example 1. We say that a random vector $\mathbf{X} \in \mathbb{R}^d$ is degenerate at a point $\mathbf{c} \in \mathbb{R}^d$ if $P(\mathbf{X} = \mathbf{c}) = 1$. Let $X_n \in \mathbb{R}^1$ be degenerate at the point $1/n$, for $n = 1, 2, \ldots$ and let $X \in \mathbb{R}^1$ be degenerate at 0. Since $1/n$ converges to zero as n tends to infinity, it may be expected that $X_n \xrightarrow{\mathscr{L}} X$. This may be seen by checking Definition 1. The distribution function of X_n is $F_{X_n}(x) = I_{[1/n, \infty)}(x)$, and that of X is $F_X(x) = I_{[0, \infty]}(x)$, where $I_A(x)$ denotes the indicator function of the set A (i.e., $I_A(x)$ denotes 1 if $x \in A$, and 0 otherwise). Then $F_{X_n}(x) \longrightarrow F_X(x)$ for all x except $x = 0$, and for $x = 0$ we have $F_{X_n}(0) = 0 \not\rightarrow F_X(0) = 1$. But because $F_X(x)$ is not continuous at $x = 0$, we nevertheless have $X_n \xrightarrow{\mathscr{L}} X$ from Definition 1. This shows the need, in the definition of convergence in law, to exclude points x at which $F_X(x)$ is not continuous.

Definition 2. $\mathbf{X}_n$ converges *in probability* to $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{X}$, if for every $\varepsilon > 0$, $P\{|\mathbf{X}_n - \mathbf{X}| > \varepsilon\} \longrightarrow 0$ as $n \longrightarrow \infty$.

Definition 3. For a real number $r > 0$, $\mathbf{X}_n$ converges *in the rth mean* to $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$, if $E|\mathbf{X}_n - \mathbf{X}|^r \longrightarrow 0$ as $n \longrightarrow \infty$.

Definition 4. $\mathbf{X}_n$ converges *almost surely* to $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$, if

$$P\{\lim_{n\to\infty} \mathbf{X}_n = \mathbf{X}\} = 1.$$

Almost sure convergence is sometimes called convergence *with probability* 1 (w.p. 1) or *strong* convergence. In statistics, convergence in the rth mean is most useful for $r = 2$, when it is called convergence *in quadratic mean*, and is written $\mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{X}$. The basic relationships are as follows.

Theorem 1.

a) $\mathbf{X}_n \xrightarrow{\text{a.s}} \mathbf{X} \Rightarrow \mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{X}$.
b) $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$ *for some* $r > 0 \Rightarrow \mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{X}$.
c) $\mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{X} \Rightarrow \mathbf{X}_n \xrightarrow{\mathscr{L}} \mathbf{X}$.

Theorem 1 states the only universally valid implications between the various modes of convergence, as the following examples show.

Example 2. To check convergence in law, nothing needs to be known about the joint distribution of $\mathbf{X}_n$ and $\mathbf{X}$, whereas this distribution must be defined to check convergence in probability. For example, if $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed (i.i.d.) normal random variables, with mean 0 and variance 1, then $X_n \xrightarrow{\mathscr{L}} X_1$, yet $X_n \overset{\mathrm{P}}{\nrightarrow} X_1$.

Example 3. Let Z be a random variable with a uniform distribution on the interval $(0, 1)$, $Z \in \mathscr{U}(0, 1)$, and let $X_1 = 1$, $X_2 = I_{[0,1/2)}(Z)$, $X_3 = iI_{[1/2,1)}(Z)$, $X_4 = I_{[0,1/4)}(Z)$, $X_5 = I_{[1/4,1/2)}(Z), \ldots$. In general, if $n = 2^k + m$, where $0 \le m < 2^k$ and $k \ge 0$, then $X_n = I_{[m2^{-k},(m+1)2^{-k})}(Z)$. Then X_n does not converge for any $Z \in [0, 1)$, so $X_n \overset{\text{a.s.}}{\nrightarrow} 0$. Yet $X_n \xrightarrow{r} 0$ for all $r > 0$ and $X_n \xrightarrow{\mathrm{P}} 0$.

Example 4. Let Z be $\mathscr{U}(0,1)$ and let $X_n = 2^n I_{[0,1/n)}(Z)$. Then $E|X_n|^r = 2^{nr}/n \longrightarrow \infty$, so $X_n \overset{r}{\not\to} 0$ for any $r > 0$. Yet $X_n \xrightarrow{\text{a.s.}} 0$ ($\{\lim_{n\to\infty} X_n = 0\} = \{Z > 0\}$, and $P\{Z > 0\} = 1$), and $X_n \xrightarrow{\text{P}} 0$ (if $0 < \varepsilon < 1$, $P(|X_n| > \varepsilon) = P(X_n = 2^n) = 1/n \longrightarrow 0$).

In this example, we have $0 \le X_n \xrightarrow{\text{a.s.}} X$ and $\lim_{n\to\infty} EX_n > EX$. That we cannot have $0 \le X_n \xrightarrow{\text{a.s.}} X$ and $\lim_{n\to\infty} EX_n < EX$ follows from the *Fatou–Lebesgue Lemma*. This states: If $X_n \xrightarrow{\text{a.s.}} X$ and if for all n $X_n \ge Y$ for some random variable Y with $E|Y| < \infty$, then $\liminf_{n\to\infty} EX_n \ge EX$. In particular, this implies the *Monotone Convergence Theorem*: If $0 \le X_1 \le X_2 \le \cdots$ and $X_n \xrightarrow{\text{a.s.}} X$, then $EX_n \longrightarrow EX$. In these theorems, X, EX_n, and EX may take the value $+\infty$.

The Fatou–Lebesgue Lemma also implies the basic *Lebesgue Dominated Convergence Theorem*: If $X_n \xrightarrow{\text{a.s.}} X$ and if $|X_n| \le Y$ for some random variable Y with $E|Y| < \infty$, then $EX_n \longrightarrow EX$.

The following lemma contains an equivalent definition of almost sure convergence. It clarifies the distinction between convergence in probability and convergence almost surely. For convergence in probability, one needs for every $\varepsilon > 0$ that the probability that $\mathbf{X}_n$ is within ε of $\mathbf{X}$ tends to one. For convergence almost surely, one needs for every $\varepsilon > 0$ that the probability that $\mathbf{X}_k$ stays within ε of $\mathbf{X}$ for all $k \ge n$ tends to one as n tends to infinity.

Lemma 1. $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ *if and only if for every* $\varepsilon > 0$,

$$P\{|\mathbf{X}_k - \mathbf{X}| < \varepsilon, \text{ for all } k \ge n\} \longrightarrow 1 \qquad \text{as } n \longrightarrow \infty. \tag{1}$$

Proof. Let $A_{n,\varepsilon} = \{|\mathbf{X}_k - \mathbf{X}| < \varepsilon \text{ for all } k \ge n\}$. Then $P\{\lim_{n\to\infty} \mathbf{X}_n = \mathbf{X}\} = P$ {for every $\varepsilon > 0$, there exists an n such that $|\mathbf{X}_k - \mathbf{X}| < \varepsilon$ for all $k \ge n$} $= P\{\bigcap_{\varepsilon>0} \bigcup_n A_{n,\varepsilon}\}$. Thus, $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ is equivalent to

$$P\left\{\bigcap_{\varepsilon>0} \bigcup_n A_{n,\varepsilon}\right\} = 1. \tag{2}$$

Because the sets $\bigcup_n A_{n,\varepsilon}$ decrease to $\bigcap_{\varepsilon>0} \bigcup_n A_{n,\varepsilon}$ as $\varepsilon \to 0$, (2) is equivalent to $P\{\bigcup_n A_{n,\varepsilon}\} = 1$ for all $\varepsilon > 0$. Then, because $A_{n,\varepsilon}$ increases to $\bigcup_n A_{n,\varepsilon}$ as $n \longrightarrow \infty$, this in turn is equivalent to

$$P\{A_{n,\varepsilon}\} \longrightarrow 1 \text{ as } n \longrightarrow \infty, \qquad \text{for all } \varepsilon > 0, \tag{3}$$

which is exactly (1). ∎

Proof of Theorem 1.

(a) $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X} \Rightarrow \mathbf{X}_n \xrightarrow{\text{P}} \mathbf{X}$: Let $\varepsilon > 0$. Then

$$P\{|\mathbf{X}_n - \mathbf{X}| \le \varepsilon\} \ge P\{|\mathbf{X}_k - \mathbf{X}| \le \varepsilon, \text{ for all } k \ge n\} \longrightarrow 1 \text{ as } n \to \infty,$$

from Lemma 1.

(b) $\mathbf{X}_n \xrightarrow{r} \mathbf{X} \Rightarrow \mathbf{X}_n \xrightarrow{\text{P}} \mathbf{X}$: We let $I(\mathbf{X} \in A)$ denote the indicator random variable that is equal to 1 if $\mathbf{X} \in A$ and to 0 otherwise. Note that

$$E|\mathbf{X}_n - \mathbf{X}|^r \ge E\big[|\mathbf{X}_n - \mathbf{X}|^r I\{|\mathbf{X}_n - \mathbf{X}| \ge \varepsilon\}\big] \ge \varepsilon^r P\{|\mathbf{X}_n - \mathbf{X}| \ge \varepsilon\}.$$

(This is *Chebyshev's Inequality*.) The result follows by letting $n \longrightarrow \infty$.

(c) $\mathbf{X}_n \xrightarrow{P} \mathbf{X} \Rightarrow \mathbf{X}_n \xrightarrow{\mathscr{L}} \mathbf{X}$: Let $\varepsilon > 0$ and let $\mathbf{1} \in \mathbb{R}^d$ represent the vector with 1 in every component. If $\mathbf{X}_n \le \mathbf{x}_0$, then either $\mathbf{X} \le \mathbf{x}_0 + \varepsilon\mathbf{1}$ or $|\mathbf{X} - \mathbf{X}_n| > \varepsilon$. In other words, $\{\mathbf{X}_n \le \mathbf{x}_0\} \subset \{\mathbf{X} \le \mathbf{x}_0 + \varepsilon\mathbf{1}\} \cup \{|\mathbf{X} - \mathbf{X}_n| > \varepsilon\}$. Hence,

$$F_{\mathbf{X}_n}(\mathbf{x}_0) \le F_{\mathbf{X}}(\mathbf{x}_0 + \varepsilon\mathbf{1}) + P\{|\mathbf{X} - \mathbf{X}_n| > \varepsilon\}.$$

Similarly,

$$F_{\mathbf{X}}(\mathbf{X}_0 - \varepsilon\mathbf{1}) \le F_{\mathbf{X}_n}(\mathbf{x}_0) + P\{|\mathbf{X} - \mathbf{X}_n| > \varepsilon\}.$$

Hence, since $P\{|\mathbf{X} - \mathbf{X}_n| > \varepsilon\} \longrightarrow 0$ as $n \longrightarrow \infty$,

$$F_{\mathbf{X}}(\mathbf{x}_0 - \varepsilon\mathbf{1}) \le \liminf F_{\mathbf{X}_n}(\mathbf{x}_0) \le \limsup \mathbf{F}_{\mathbf{X}_n}(\mathbf{x}_0) \le F_{\mathbf{X}}(\mathbf{x}_0 + \varepsilon\mathbf{1}).$$

If $F_{\mathbf{X}}(\mathbf{x})$ is continuous at $\mathbf{x}_0$, then the left and right ends of this inequality both converge to $F_{\mathbf{X}}(\mathbf{x}_0)$ as $\varepsilon \longrightarrow 0$, implying that

$$F_{\mathbf{X}_n}(\mathbf{x}_0) \longrightarrow F_{\mathbf{X}}(\mathbf{x}_0) \text{ as } n \longrightarrow \infty. \quad \blacksquare$$

EXERCISES

1. Suppose $X_n \in \mathscr{B}e(1/n, 1/n)$ (beta) and $X \in \mathscr{B}(1, 1/2)$ (binomial). Show that $X_n \xrightarrow{\mathscr{L}} X$. What if $X_n \in \mathscr{B}e(\alpha/n, \beta/n)$?
2. Suppose X_n is uniformly distributed on the set of points $\{1/n, 2/n, \dots, 1\}$. Show that $X_n \xrightarrow{\mathscr{L}} X$, where X is $\mathscr{U}(0,1)$. Does $X_n \xrightarrow{\text{P}} X$?
3. (a) Show that if $0 < r' < r$ and $E|X|^r < \infty$, then $E|X|^{r'} < \infty$.
   (b) Show that if $0 < r' < r$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{r'} X$. You may use *Hölder's Inequality*: For nonnegative random variables X and Y with finite means, $EX^pY^{1-p} \le (EX)^p(EY)^{1-p}$ for $0 \le p \le 1$.
4. Give an example of random variables X_n such that $E|X_n| \to 0$ and $E|X_n|^2 \longrightarrow 1$.

5. Let μ be a constant. Show that $X_n \xrightarrow{\text{qm}} \mu$ if and only if $EX_n \to \mu$, and $\text{var}(X_n) \to 0$.
6. If the limiting distribution function, $F_{\mathbf{X}}$, is continuous, then the definition of convergence in law is simply that $F_{\mathbf{X}_n}(\mathbf{x}) \to F_{\mathbf{X}}(\mathbf{x})$ as $n \to \infty$, for all $\mathbf{x}$. However, in this case, it automatically follows that the convergence is uniform in $\mathbf{x}$. Prove this in one dimension: If F_X is continuous and $X_n \xrightarrow{\mathscr{L}} X$ as $n \to \infty$, then $\sup_x |F_{X_n}(x) - F_X(x)| \to 0$ as $n \to \infty$.
7. Using the Fatou–Lebesgue Lemma, (a) prove the Monotone Convergence Theorem, and (b) prove the Lebesgue Dominated Convergence Theorem.

2

Partial Converses to Theorem 1

Although complete converses to the statements of Theorem 1 are invalid, as we have seen, under certain additional conditions some important partial converses hold. We use the same symbol **c** to denote the point $\mathbf{c} \in \mathbb{R}^d$, as well as the degenerate random vector identically equal to **c**.

Theorem 2.

(a) *If* $\mathbf{c} \in \mathbb{R}^d$, *then* $\mathbf{X}_n \xrightarrow{\mathscr{L}} \mathbf{c} \Rightarrow \mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{c}$.
(b) *If* $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ *and* $|\mathbf{X}_n|^r \leq Z$ *for some* $r > 0$ *and some random variable* Z *with* $EZ < \infty$, *then* $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$.
(c) [*Scheffé* (1947)]. *If* $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$, $\mathbf{X}_n \geq 0$, *and* $E\mathbf{X}_n \to E\mathbf{X} < \infty$, *then* $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$, *where* $r = 1$.
(d) $\mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{X}$ *if and only if every subsequence* $n_1, n_2, \ldots \varepsilon \{1, 2, \ldots\}$ *has a sub-sequence* $m_1, m_2, \ldots \varepsilon \{n_1, n_2, \ldots\}$ *such that* $\mathbf{X}_{m_j} \xrightarrow{\text{a.s.}} \mathbf{X}$ *as* $j \to \infty$.

Remarks. Part (a), together with part (c) of Theorem 1, implies that convergence in law and convergence in probability are equivalent if the limit is a constant random vector. In the following sections we use this equivalence often without explicit mention.

Part (b) gives a method of deducing convergence in the rth mean from almost sure convergence. See Exercise 3 for a strengthening of this result, and Exercise 2 for a simple sufficient condition for almost sure convergence.

Part (c) is sometimes called Scheffé's Useful Convergence Theorem because of the title of Scheffé's 1947 article. It is usually stated in terms of densities (nonnegative functions that integrate to one) as follows: If $f_n(x)$ and $g(x)$ are densities such that $f_n(x) \to g(x)$ for all x, then $\int |f_n(x) - g(x)|\, dx \to 0$. [The hypotheses $f_n(x) \geq 0$ and $\int f_n(x)\, dx \to \int g(x)\, dx$ are automatic here. The proof of this is analogous to the proof of (c) given below.]

Pointwise convergence of densities is a type of convergence in distribution that is much stronger than convergence in law. Convergence in law only requires that $P(\mathbf{X}_n \in A)$ converge to $P(\mathbf{X} \in A)$ for certain sets A of the form $\{\mathbf{x}: \mathbf{x} \le \mathbf{a}\}$. If the densities converge, then $P(\mathbf{X}_n \in A)$ converges to $P(\mathbf{X} \in A)$ for all Borel sets A, and, moreover, the convergence is uniform in A. In other words, suppose that $\mathbf{X}_n$ and $\mathbf{X}$ have densities (with respect to a measure ν) denoted by $f_n(\mathbf{x})$ and $f(\mathbf{x})$, respectively. Then, if $f_n(\mathbf{x}) \longrightarrow f(\mathbf{x})$ for all $\mathbf{x}$, we have

$$\sup_A |P(\mathbf{X}_n \in A) - P(\mathbf{X} \in A)| \longrightarrow 0.$$

The proof is an exercise. We will encounter this type of convergence later in the Bernstein–von Mises Theorem.

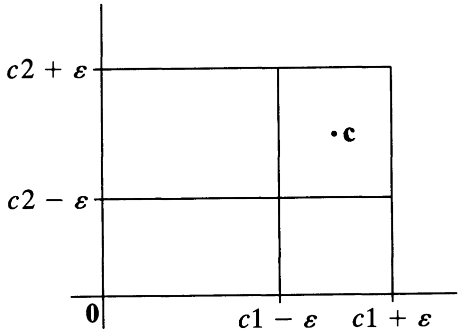
As an illustration of the difference between this type of convergence and convergence in law, suppose that X_n is uniformly distributed on the set $\{1/n, 2/n, \ldots, n/n\}$. Then $X_n \xrightarrow{\mathscr{L}} X \in \mathscr{U}(0,1)$, the uniform distribution on $[0,1]$, but $P(\mathbf{X}_n \in A)$ does not converge to $P(\mathbf{X} \in A)$ for all A. For example, if $A = \{x\colon x \text{ is rational}\}$, then $P(\mathbf{X}_n \in A) = 1$ does not converge to $P(\mathbf{X} \in A) = 0$.

Part (d) is a tool for dealing with convergence in probability using convergence almost surely. Generally convergence almost surely is easier to work with. Here is an example of the use of part (d). If $\mathbf{X}_n \longrightarrow \mathbf{X}$ with probability one (i.e., almost surely), and if $g(\mathbf{x})$ is a continuous function of $\mathbf{x}$, then it is immediate that $g(\mathbf{X}_n) \longrightarrow g(\mathbf{X})$ with probability one. Is the same result true if convergence almost surely is replaced by convergence in probability? Assume $\mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{X}$ and let $g(\mathbf{x})$ be a continuous function of $\mathbf{x}$. To show $g(\mathbf{X}_n) \xrightarrow{\mathrm{P}} g(\mathbf{X})$, it is sufficient, according to part (d), to show that for every subsequence, $n_1, n_2, \ldots \varepsilon \{1, 2, \ldots\}$, there is a sub-subsequence, $m_1, m_2, \ldots \varepsilon \{n_1, n_2, \ldots\}$ such that $g(\mathbf{X}_{m_i}) \xrightarrow{\text{a.s.}} g(\mathbf{X})$ as $i \longrightarrow \infty$. So let $n_1, n_2, \ldots$ be an arbitrary subsequence and find, using part (d), a sub-subsequence $m_1, m_2, \ldots \varepsilon \{n_1, n_2, \ldots\}$ so that $\mathbf{X}_{m_i} \xrightarrow{\text{a.s.}} \mathbf{X}$. Then $g(\mathbf{X}_{m_i}) \xrightarrow{\text{a.s.}} g(\mathbf{X})$, since $g(\mathbf{x})$ is continuous, and the result is proved.

Proof of Theorem 2. (a) (In two dimensions)

$$\begin{aligned}
P\{|\mathbf{X}_n - \mathbf{c}| \le \varepsilon\sqrt{2}\} &\ge P\left\{\mathbf{c} - \varepsilon\begin{pmatrix}1\\1\end{pmatrix} < \mathbf{X}_n \le \mathbf{c} + \varepsilon\begin{pmatrix}1\\1\end{pmatrix}\right\} \\
&= P\left\{\mathbf{X}_n \le \mathbf{c} + \varepsilon\begin{pmatrix}1\\1\end{pmatrix}\right\} - P\left\{\mathbf{X}_n \le \mathbf{c} + \varepsilon\begin{pmatrix}1\\-1\end{pmatrix}\right\} \\
&\quad - P\left\{\mathbf{X}_n \le \mathbf{c} + \varepsilon\begin{pmatrix}-1\\1\end{pmatrix}\right\} + P\left\{\mathbf{X}_n \le \mathbf{c} - \varepsilon\begin{pmatrix}1\\1\end{pmatrix}\right\}
\end{aligned}$$

Here is a picture:



(b) This is the Lebesgue Dominated Convergence Theorem in d dimensions. Note that $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ and $|\mathbf{X}_n|^r \le Z$ implies $|\mathbf{X}|^r \le Z$ a.s., so that $|\mathbf{X}_n - \mathbf{X}|^r \le (|\mathbf{X}_n| + |\mathbf{X}|)^r \le (Z^{1/r} + Z^{1/r})^r \le 2^r Z$ a.s. Now apply the Lebesgue Dominated Convergence Theorem in the form given in the previous section replacing X_n by $|\mathbf{X}_n - \mathbf{X}|^r$ and X by 0.

(c) Let x^+ denote the positive part of x: $x^+ = \max\{0, x\}$. In one dimension, for a real number x, $|x| = x + 2(-x)^+$; hence $E|X_n - X| = E(X_n - X) + 2E(X - X_n)^+$. The first term converges to zero because $EX_n \to EX$. The second term converges to zero by the Lebesgue Dominated Convergence Theorem, because $0 \le (X - X_n)^+ \le X^+$ and $EX^+ < \infty$. For dimensions greater than one, use the triangle inequality, $|\mathbf{X}_n - \mathbf{X}| \le \sum_{j=1}^d |X_{nj} - X_j|$, and use the above analysis on each term separately.

The proof of part (d) is based on the Borel–Cantelli Lemma. For events A_j, $j = 0, 1, \ldots$, the event $\{A_j \text{ i.o.}\}$ (read A_j *infinitely often*), stands for the event that infinitely many A_j occur.

The Borel–Cantelli Lemma. *If $\sum_{j=1}^\infty P(A_j) < \infty$, then $P\{A_j \text{ i.o.}\} = 0$. Conversely, if the A_j are independent and $\sum_{j=1}^\infty P(A_j) = \infty$, then $P\{A_j \text{ i.o.}\} = 1$.*

Proof. (The general half) If infinitely many of the A_j occur, then for all n, at least one A_j with $j \ge n$ occurs. Hence,

$$P\{A_j \text{ i.o.}\} \le P\left\{\bigcup_{j=n}^\infty A_j\right\} \le \sum_{j=n}^\infty P(A_j) \to 0. \quad \blacksquare$$

The proof of the converse is an exercise. (See Exercise 4.)

A typical example of the use of the Borel–Cantelli Lemma occurs in coin tossing. Let $X_1, X_2, \ldots$ be a sequence of independent Bernoulli trials with probability of success on the nth trial equal to p_n. What is the probability of an infinite number of successes? Or, equivalently, what is $P\{X_n = 1 \text{ i.o.}\}$? From the Borel–Cantelli Lemma and its converse, this probability is zero or one depending on whether $\sum p_n < \infty$ or not. If $p_n = 1/n^2$, for example, then $P\{X_n = 1 \text{ i.o.}\} = 0$. If $p_n = 1/n$, then $P\{X_n = 1 \text{ i.o.}\} = 1$.

The Borel–Cantelli Lemma is useful in dealing with problems involving almost sure convergence because $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ is equivalent to

$$P\{|\mathbf{X}_n - \mathbf{X}| > \varepsilon \text{ i.o.}\} = 0, \qquad \text{for all } \varepsilon > 0.$$

(d) (If) Suppose $\mathbf{X}_n$ does not converge in probability to $\mathbf{X}$. Then there exists an $\varepsilon > 0$ and a $\delta > 0$ such that $P\{|X_n - X| > \varepsilon\} > \delta$ for infinitely many n, say $\{n_j\}$. Then no subsequence of $\{n_j\}$ converges in probability, nor, consequently, almost surely.

(Only if) Let $\varepsilon_n > 0$ and $\sum_{j=1}^{\infty} \varepsilon_j < \infty$. Find n_j such that $P\{|\mathbf{X}_n - \mathbf{X}| \geq \varepsilon_j\} < \varepsilon_j$ for all $n \geq n_j$, and assume without loss of generality that $n_1 < n_2 < \ldots$. Let $A_j = \{|\mathbf{X}_{n_j} - \mathbf{X}| \geq \varepsilon_j\}$. Then, $\sum_{j=1}^{\infty} P(A_j) \leq \sum_{j=1}^{\infty} \varepsilon_j < \infty$, so by the Borel–Cantelli Lemma, $P\{A_j \text{ i.o.}\} = 0$. This says that with probability 1, $|\mathbf{X}_{n_j} - \mathbf{X}| \geq \varepsilon_j$ occurs only finitely many times. Since $\varepsilon_j \to 0$, we have for any $\varepsilon > 0$ that with probability 1, $|\mathbf{X}_{n_j} - \mathbf{X}| \geq \varepsilon$ occurs only finitely many times. Hence, $\mathbf{X}_{n_j} \xrightarrow{\text{a.s.}} \mathbf{X}$; that is $P\{|\mathbf{X}_{n_j} - \mathbf{X}| > \varepsilon \text{ i.o.}\} = 0$ for all $\varepsilon > 0$. Similarly, if n' is any subsequence, $\mathbf{X}_{n'} \xrightarrow{\text{P}} \mathbf{X}$, so we can find a sub-subsequence n'' of n' such that $\mathbf{X}_{n''} \xrightarrow{\text{a.s.}} \mathbf{X}$. ■

EXERCISES

1. Let $X_1, X_2, \ldots$ be independent identically distributed with densities $f(x) = \alpha x^{-(\alpha+1)} I_{(1,\infty)}(x)$. (a) For what values of $\alpha > 0$ and $r > 0$ is it true that $(1/n)X_n \xrightarrow{r} 0$? (b) For what values of $\alpha > 0$ is it true that $(1/n)X_n \xrightarrow{\text{a.s.}} 0$? (Use the Borel–Cantelli Lemma.)
2. Show that if $\sum E(\mathbf{X}_n - \mathbf{X})^2 < \infty$, then $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ and $\mathbf{X}_n \xrightarrow{\text{qm}} \mathbf{X}$. Show that if $\sum E|\mathbf{X}_n - \mathbf{X}|^r < \infty$, then $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ and $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$.
3. Improve Theorem 2(b) and Theorem 2(c) by using Theorem 2(d) to show
   (a) If $\mathbf{X}_n \xrightarrow{\text{P}} \mathbf{X}$ and $|\mathbf{X}_n|^r \leq Z$ for some $r > 0$ and some random variable $\mathbf{Z}$ such that $EZ < \infty$, then $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$.
   (b) If $\mathbf{X}_n \xrightarrow{\text{P}} \mathbf{X}$, $\mathbf{X}_n \geq 0$, and $E\mathbf{X}_n \to E\mathbf{X} < \infty$, then $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$, where $r = 1$.