Probability and Its Applications

Published in association with the Applied Probability Trust

Editors: J. Gani, C.C. Heyde, P. Jagers, T.G. Kurtz

Probability and Its Applications

Azencott et al.: Series of Irregular Observations. Forecasting and Model Building. 1986
Bass: Diffusions and Elliptic Operators. 1997
Bass: Probabilistic Techniques in Analysis. 1995
Berglund/Gentz: Noise-Induced Phenomena in Slow-Fast Dynamical Systems: A Sample-Paths Approach. 2006
Biagini/Hu/Øksendal/Zhang: Stochastic Calculus for Fractional Brownian Motion and Applications. 2008
Chen: Eigenvalues, Inequalities and Ergodic Theory. 2005
Costa/Fragoso/Marques: Discrete-Time Markov Jump Linear Systems. 2005
Daley/Vere-Jones: An Introduction to the Theory of Point Processes I: Elementary Theory and Methods. 2nd ed. 2003, corr. 2nd printing 2005
Daley/Vere-Jones: An Introduction to the Theory of Point Processes II: General Theory and Structure. 2nd ed. 2008
de la Peña/Gine: Decoupling: From Dependence to Independence, Randomly Stopped Processes U-Statistics and Processes Martingales and Beyond. 1999
de la Peña/Lai/Shao: Self-Normalized Processes. 2009
Del Moral: Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications. 2004
Durrett: Probability Models for DNA Sequence Evolution. 2002, 2nd ed. 2008
Galambos/Simonelli: Bonferroni-Type Inequalities with Equations. 1996
Gani (ed.): The Craft of Probabilistic Modelling. A Collection of Personal Accounts. 1986
Gut: Stopped Random Walks. Limit Theorems and Applications. 1987
Guyon: Random Fields on a Network. Modeling, Statistics and Applications. 1995
Kallenberg: Foundations of Modern Probability. 1997, 2nd ed. 2002
Kallenberg: Probabilistic Symmetries and Invariance Principles. 2005
Last/Brandt: Marked Point Processes on the Real Line. 1995
Molchanov: Theory of Random Sets. 2005
Nualart: The Malliavin Calculus and Related Topics, 1995, 2nd ed. 2006
Schmidli: Stochastic Control in Insurance. 2008
Schneider/Weil: Stochastic and Integral Geometry. 2008
Shedler: Regeneration and Networks of Queues. 1986
Silvestrov: Limit Theorems for Randomly Stopped Stochastic Processes. 2004
Rachev/Rueschendorf: Mass Transportation Problems. Volume I: Theory and Volume II: Applications. 1998
Resnick: Extreme Values, Regular Variation and Point Processes. 1987
Thorisson: Coupling, Stationarity and Regeneration. 2000

Victor H. de la Peña · Tze Leung Lai · Qi-Man Shao

Self-Normalized Processes

Limit Theory and Statistical Applications

Springer

Victor H. de la Peña
Department of Statistics
Columbia University
Mail Code 4403
New York, NY 10027
USA
vp@stat.columbia.edu

Tze Leung Lai
Department of Statistics
Sequoia Hall, 390 Serra Mall
Stanford University
Stanford, CA 94305-4065
USA
lait@stat.stanford.edu

Qi-Man Shao
Department of Mathematics
Hong Kong University of Science and Technology
Clear Water Bay
Kowloon, Hong Kong
People's Republic of China
maqmshao@ust.hk

Series Editors:

ISBN: 978-3-540-85635-1 e-ISBN: 978-3-540-85636-8

Probability and Its Applications ISSN print edition: 1431-7028

Library of Congress Control Number: 2008938080

Mathematics Subject Classification (2000): Primary: 60F10, 60F15, 60G50, 62E20;
Secondary: 60E15, 60G42, 60G44, 60G40, 62L10

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To our families

for V.H.P., Colleen, Victor, Mary-Margaret and Patrick

for T.L.L., Letitia, Peter and David

for Q.-M.S., Jiena and Wenqi

Preface

This year marks the centennial of Student's seminal 1908 paper, "On the probable error of a mean," in which the t-statistic and the t-distribution were introduced. During the past century, the t-statistic has evolved into much more general Studentized statistics and self-normalized processes, and the t-distribution generalized to the multivariate case, leading to multivariate processes with matrix self-normalization and bootstrap-t methods for tests and confidence intervals. The past two decades have also witnessed the active development of a rich probability theory of self-normalized processes, beginning with laws of the iterated logarithm, weak convergence, large and moderate deviations for self-normalized sums of independent random variables, and culminating in exponential and moment bounds and a universal law of the iterated logarithm for self-normalized processes in the case of dependent random variables. An important goal of this book is to present the main techniques and results of these developments in probability and to relate them to the asymptotic theory of Studentized statistics and to other statistical applications.

Another objective of writing this book is to use it as course material for a Ph.D. level course on selected topics in probability theory and its applications. Lai and Shao co-taught such a course for Ph.D. students in the Department of Statistics at Stanford University in the summer of 2007. These students had taken the Ph.D. core courses in probability (at the level of Durrett's *Probability: Theory and Examples*) and in theoretical statistics (at the level of Lehmann's *Testing Statistical Hypotheses* and *Theory of Point Estimation*). They found the theory of self-normalized processes an attractive topic, supplementing and integrating what they had learned from their core courses in probability and theoretical statistics and also exposing them to new techniques and research topics in both areas. The success of the experimental course STATS 300 (Advanced Topics in Statistics and Probability) prompted Lai and Shao to continue offering it periodically at Stanford and Hong Kong University of Science and Technology. A similar course is being planned at Columbia University by de la Peña. With these courses in mind, we have included exercises and supplements for the reader to explore related concepts and methods not covered in introductory Ph.D.-level courses, besides providing basic references related to these topics. We also plan to update these periodically at the Web site for the book: `http://www.math.ust.hk/~maqmshao/book-self/SNP.html`.

We acknowledge grant support for our research projects related to this book from the National Science Foundation (DMS-0505949 and 0305749) and the Hong Kong Research Grants Council (CERG-602206 and 602608). We thank three anonymous reviewers for their valuable suggestions, and all the students who took STATS 300 for their interest in the subject and comments on an earlier draft of certain chapters of the book that were used as lecture notes. We also thank our collaborators Hock Peng Chan, Bing-Yi Jing, Michael Klass, David Siegmund, Qiying Wang and Wang Zhou for working with us on related projects and for their helpful comments. We are particularly grateful to Cindy Kirby who helped us to coordinate our writing efforts and put together the separate chapters in an efficient and timely fashion. Without her help, this book would not have been completed in 2008 to commemorate Student's centennial.

Department of Statistics, Columbia University *Victor H. de la Peña*
Department of Statistics, Stanford University *Tze Leung Lai*
Department of Mathematics, Hong Kong University of Science & Technology *Qi-Man Shao*

Contents

1 Introduction 1

Part I Independent Random Variables

2 Classical Limit Theorems, Inequalities and Other Tools 7
- 2.1 Classical Limit Theorems 7
  - 2.1.1 The Weak Law, Strong Law and Law of the Iterated Logarithm 8
  - 2.1.2 The Central Limit Theorem 9
  - 2.1.3 Cramér's Moderate Deviation Theorem 11
- 2.2 Exponential Inequalities for Sample Sums 11
  - 2.2.1 Self-Normalized Sums 11
  - 2.2.2 Tail Probabilities for Partial Sums 13
- 2.3 Characteristic Functions and Expansions Related to the CLT 17
  - 2.3.1 Continuity Theorem and Weak Convergence 18
  - 2.3.2 Smoothing, Local Limit Theorems and Expansions 19
- 2.4 Supplementary Results and Problems 21

3 Self-Normalized Large Deviations 25
- 3.1 A Classical Large Deviation Theorem for Sample Sums 25
- 3.2 A Large Deviation Theorem for Self-Normalized Sums 27
  - 3.2.1 Representation by Supremum over Linear Functions of (S_n, V_n^2) 27
  - 3.2.2 Proof of Theorem 3.1 28
- 3.3 Supplementary Results and Problems 31

4 Weak Convergence of Self-Normalized Sums 33
- 4.1 Self-Normalized Central Limit Theorem 33
- 4.2 Non-Normal Limiting Distributions for Self-Normalized Sums 37
- 4.3 Supplementary Results and Problems 38

5 Stein's Method and Self-Normalized Berry–Esseen Inequality 41
5.1 Stein's Method 41
5.1.1 The Stein Equation 41
5.1.2 Stein's Method: Illustration of Main Ideas 44
5.1.3 Normal Approximation for Smooth Functions 46
5.2 Concentration Inequality and Classical Berry–Esseen Bound 49
5.3 A Self-Normalized Berry–Esseen Inequality 52
5.3.1 Proof: Outline of Main Ideas 53
5.3.2 Proof: Details 55
5.4 Supplementary Results and Problems 60

6 Self-Normalized Moderate Deviations and Laws of the Iterated Logarithm 63
6.1 Self-Normalized Moderate Deviations: Normal Case 63
6.1.1 Proof of the Upper Bound 64
6.1.2 Proof of the Lower Bound 66
6.2 Self-Normalized Moderate Deviations: Stable Case 69
6.2.1 Preliminary Lemmas 70
6.2.2 Proof of Theorem 6.6 76
6.3 Self-Normalized Laws of the Iterated Logarithm 81
6.4 Supplementary Results and Problems 84

7 Cramér-Type Moderate Deviations for Self-Normalized Sums 87
7.1 Self-Normalized Cramér-Type Moderate Deviations 87
7.2 Proof of Theorems 90
7.2.1 Proof of Theorems 7.2, 7.4 and Corollaries 90
7.2.2 Proof of Theorem 7.1 91
7.2.3 Proof of Propositions 94
7.3 Application to Self-Normalized LIL 96
7.4 Cramér-Type Moderate Deviations for Two-Sample t-Statistics 104
7.5 Supplementary Results and Problems 106

8 Self-Normalized Empirical Processes and U-Statistics 107
8.1 Self-Normalized Empirical Processes 107
8.2 Self-Normalized U-Statistics 108
8.2.1 The Hoeffding Decomposition and Central Limit Theorem 109
8.2.2 Self-Normalized U-Statistics and Berry–Esseen Bounds 109
8.2.3 Moderate Deviations for Self-Normalized U-Statistics 110
8.3 Proofs of Theorems 8.5 and 8.6 111
8.3.1 Main Ideas of the Proof 111
8.3.2 Proof of Theorem 8.6 112
8.3.3 Proof of Theorem 8.5 113
8.3.4 Proof of Proposition 8.7 113
8.4 Supplementary Results and Problems 119

Part II Martingales and Dependent Random Vectors

9 Martingale Inequalities and Related Tools 123
9.1 Basic Martingale Theory 123
9.1.1 Conditional Expectations and Martingales 123
9.1.2 Martingale Convergence and Inequalities 125
9.2 Tangent Sequences and Decoupling Inequalities 125
9.2.1 Construction of Decoupled Tangent Sequences 126
9.2.2 Exponential Decoupling Inequalities 126
9.3 Exponential Inequalities for Martingales 128
9.3.1 Exponential Inequalities via Decoupling 128
9.3.2 Conditionally Symmetric Random Variables 132
9.3.3 Exponential Supermartingales and Associated Inequalities . . 134
9.4 Supplementary Results and Problems 135

10 A General Framework for Self-Normalization 137
10.1 An Exponential Family of Supermartingales Associated with Self-Normalization 137
10.1.1 The I.I.D. Case and Another Derivation of *(3.8)* 137
10.1.2 A Representation of Self-Normalized Processes and Associated Exponential Supermartingales 138
10.2 Canonical Assumptions and Related Stochastic Models 139
10.3 Continuous-Time Martingale Theory 140
10.3.1 Doob–Meyer Decomposition and Locally Square-Integrable Martingales 141
10.3.2 Inequalities and Stochastic Integrals 143
10.4 Supplementary Results and Problems 146

11 Pseudo-Maximization via Method of Mixtures 149
11.1 Pseudo-Maximization and Laplace's Method 149
11.2 A Class of Mixing Densities 150
11.3 Application of Method of Mixtures to Boundary Crossing Probabilities 152
11.3.1 The Robbins–Siegmund Boundaries for Brownian Motion . . 152
11.3.2 Extensions to General Self-Normalized Processes 154
11.4 Supplementary Results and Problems 157

12 Moment and Exponential Inequalities for Self-Normalized Processes 161
12.1 Inequalities of Caballero, Fernandez and Nualart, Graversen and Peskir, and Kikuchi 161
12.2 Moment Bounds via the Method of Mixtures 164
12.2.1 Gaussian Mixing Densities 165
12.2.2 The Mixing Density Functions in Sect. 11.2 167
12.3 Applications and Examples 174
12.3.1 Proof of Lemma 8.11 174
12.3.2 Generalizations of Theorems 12.1, 12.2 and 12.3 175

12.3.3 Moment Inequalities Under Canonical Assumption for a Restricted Range . . . 176
12.4 Supplementary Results and Problems . . . 177

13 Laws of the Iterated Logarithm for Self-Normalized Processes . . . 179
13.1 Stout's LIL for Self-Normalized Martingales . . . 179
13.2 A Universal Upper LIL . . . 182
13.3 Compact LIL for Self-Normalized Martingales . . . 186
13.4 Supplementary Results and Problems . . . 190

14 Multivariate Self-Normalized Processes with Matrix Normalization . . 193
14.1 Multivariate Extension of Canonical Assumptions . . . 193
14.1.1 Matrix Sequence Roots for Self-Normalization . . . 193
14.1.2 Canonical Assumptions for Matrix-Normalized Processes . . 194
14.2 Moment and Exponential Inequalities via Pseudo-Maximization . . . 196
14.3 LIL and Boundary Crossing Probabilities for Multivariate Self-Normalized Processes . . . 201
14.4 Supplementary Results and Problems . . . 202

Part III Statistical Applications

15 The *t*-Statistic and Studentized Statistics . . . 207
15.1 Distribution Theory of Student's t-Statistics . . . 207
15.1.1 Case of Infinite Second Moment . . . 208
15.1.2 Saddlepoint Approximations . . . 210
15.1.3 The t-Test and a Sequential Extension . . . 212
15.2 Multivariate Extension and Hotelling's T^2-Statistic . . . 213
15.2.1 Sample Covariance Matrix and Wishart Distribution . . . 213
15.2.2 The Multivariate t-Distribution and Hotelling's T^2-Statistic . . . 213
15.2.3 Asymptotic Theory in the Case of Non-Normal Y_i . . . 215
15.3 General Studentized Statistics . . . 216
15.3.1 Martingale Central Limit Theorems and Asymptotic Normality . . . 216
15.3.2 Non-Normal Limiting Distributions in Unit-Root Nonstationary Autoregressive Models . . . 217
15.3.3 Studentized Statistics in Stochastic Regression Models . . . 218
15.4 Supplementary Results and Problems . . . 221

16 Self-Normalization for Approximate Pivots in Bootstrapping . . . 223
16.1 Approximate Pivots and Bootstrap-t Confidence Intervals . . . 223
16.2 Edgeworth Expansions and Second-Order Accuracy . . . 224
16.2.1 Edgeworth Expansions for Smooth Functions of Sample Means . . . 224
16.2.2 Edgeworth and Cornish–Fisher Expansions: Applications to Bootstrap-t and Percentile Intervals . . . 225

16.3 Asymptotic *U*-Statistics and Their Bootstrap Distributions 228
16.4 Application of Cramér-Type Moderate Deviations 232
16.5 Supplementary Results and Problems 233

17 Pseudo-Maximization in Likelihood and Bayesian Inference 235
17.1 Generalized Likelihood Ratio Statistics 235
17.1.1 The Wilks and Wald Statistics 236
17.1.2 Score Statistics and Their Martingale Properties 238
17.2 Penalized Likelihood and Bayesian Inference 238
17.2.1 Schwarz's Bayesian Selection Criterion 239
17.2.2 Pseudo-Maximization and Frequentist Properties of Bayes Procedures 240
17.3 Supplementary Results and Problems 241

18 Sequential Analysis and Boundary Crossing Probabilities for Self-Normalized Statistics 243
18.1 Information Bounds and Asymptotic Optimality of Sequential GLR Tests 244
18.1.1 Likelihood Ratio Identities, the Wald–Hoeffding Lower Bounds and their Asymptotic Generalizations 244
18.1.2 Asymptotic Optimality of 2-SPRTs and Sequential GLR Tests 247
18.2 Asymptotic Approximations via Method of Mixtures and Geometric Integration 251
18.2.1 Boundary Crossing Probabilities for GLR Statistics via Method of Mixtures 251
18.2.2 A More General Approach Using Saddlepoint Approximations and Geometric Integration 252
18.2.3 Applications and Examples 257
18.3 Efficient Monte Carlo Evaluation of Boundary Crossing Probabilities 260
18.4 Supplementary Results and Problems 262

References 267

Index 273

Chapter 1
Introduction

A prototypical example of a self-normalized process is Student's t-statistic based on a sample of normal i.i.d. observations $X_1, \dots, X_n$, dating back to 1908 when William Gosset ("Student") considered the problem of statistical inference on the mean μ when the standard deviation σ of the underlying distribution is unknown. Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ be the sample mean and $s_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ be the sample variance. Gosset (1908) derived the distribution of the t-statistic $T_n = \sqrt{n}(\bar{X}_n - \mu)/s_n$ for normal X_i; this is the t-distribution with $n-1$ degrees of freedom. The t-distribution converges to the standard normal distribution, and in fact T_n has a limiting standard normal distribution as $n \to \infty$ even when the X_i are nonnormal. When nonparametric methods were subsequently introduced, the t-test was compared with the nonparametric tests (e.g., the sign test and rank tests), in particular for "fat-tailed" distributions with infinite second or even first absolute moments. It has been found that the t-test of $\mu = \mu_0$ is robust against non-normality in terms of the Type I error probability but not the Type II error probability. Without loss of generality, consider the case $\mu_0 = 0$ so that

$$T_n = \frac{\sqrt{n}\bar{X}_n}{s_n} = \frac{S_n}{V_n} \left\{ \frac{n-1}{n - (S_n/V_n)^2} \right\}^{1/2}, \tag{1.1}$$

where $S_n = \sum_{i=1}^n X_i, V_n^2 = \sum_{i=1}^n X_i^2$. Efron (1969) and Logan et al. (1973) have derived limiting distributions of self-normalized sums S_n/V_n. In view of (1.1), if T_n or S_n/V_n has a limiting distribution, then so does the other, and it is well known that they coincide; see, e.g., Proposition 1 of Griffin (2002).

Active development of the probability theory of self-normalized processes began in the 1990s with the seminal work of Griffin and Kuelbs (1989, 1991) on laws of the iterated logarithm for self-normalized sums of i.i.d. variables belonging to the domain of attraction of a normal or stable law. Subsequently, Bentkus and Götze (1996) derived a Berry–Esseen bound for Student's t-statistic, and Giné et al. (1997) proved that the t-statistic has a limiting standard normal distribution if and only if X_i is in the domain of attraction of a normal law. Moreover, Csörgő et al. (2003a)

V.H. de la Peña et al., *Self-Normalized Processes: Limit Theory and Statistical Applications*, Probability and its Applications,

proved a self-normalized version of the weak invariance principle under the same necessary and sufficient condition. Shao (1997) proved large deviation results for S_n/V_n without moment conditions and moderate deviation results when X_i is the domain of attraction of a normal or stable law. Subsequently Shao (1999) obtained Cramér-type large deviation results when $E|X_1|^3 < \infty$. Jing et al. (2004) derived saddlepoint approximations for Student's t-statistic with no moment assumptions. Bercu et al. (2002) obtained large and moderate deviation results for self-normalized empirical processes. Self-normalized sums of independent but non-identically distributed X_i have been considered by Bentkus et al. (1996), Wang and Jing (1999), Jing et al. (2003) and Csörgő et al. (2003a).

Part I of the book presents in Chaps. 3–7 the basic ideas and results in the probability theory of self-normalized sums of independent random variables described above. It also extends in Chap. 8 the theory to self-normalized U-statistics based on independent random variables. Part II considers self-normalized processes in the case of dependent variables. Like Part I that begins by introducing some basic probability theory for sums of independent random variables in Chap. 2, Part II begins by giving in Chap. 9 an overview of martingale inequalities and related results which will be used in the subsequent chapters. Chapter 10 provides a general framework for self-normalization, which links the approach of de la Peña et al. (2000, 2004) for general self-normalized processes to that of Shao (1997) for large deviations of self-normalized sums of i.i.d. random variables. This general framework is also applicable to dependent random vectors that involve matrix normalization, as in Hotelling's T^2-statistic which generalizes Student's t-statistic to the multivariate case. In particular, it is noted in Chap. 10 that a basic ingredient in Shao's (1997) self-normalized large deviations theory is $e^{\psi(\theta,\rho)} := E\exp\{\theta X_1 - \rho\theta^2 X_1^2\}$, which is always finite for $\rho > 0$. This can be readily extended to the multivariate case by replacing θX_1 with $\theta' X_1$, where θ and X_1 are d-dimensional vectors. Under the assumptions $EX_1 = 0$ and $E\|X_1\|^2 < \infty$, Taylor's theorem yields

$$\psi(\theta,\rho) = \log\left(E\exp\left\{\theta' X_1 - \rho(\theta' X_1)^2\right\}\right) = \left\{\left(\frac{1}{2} - \rho + o(1)\right)\theta' E(X_1 X_1')\theta\right\}$$

as $\theta \to 0$. Let $\gamma > 0, C_n = (1+\gamma)\Sigma_{i=1}^n X_i X_i', A_n = \Sigma_{i=1}^n X_i$. It then follows that ρ and ε can be chosen sufficiently small so that

$$\begin{gathered}\left\{\exp(\theta' A_n - \theta' C_n \theta/2),\ \ \mathscr{F}_n, n \geq 1\right\} \\ \text{is a supermartingale with mean } \leq 1, \text{ for } \|\theta\| < \varepsilon.\end{gathered} \tag{1.2}$$

Note that (1.2) implies that $\{\int_{\|\theta\|<\varepsilon} e^{\theta' A_n - \theta' C_n \theta/2} f(\theta) d\theta,\ \mathscr{F}_n, n \geq 1\}$ is also a supermartingale, for any probability density f on the ball $\{\theta : \|\theta\| < \varepsilon\}$.

In Chap. 11 and its multivariate extension given in Chap. 14, we show that the supermartingale property (1.2), its weaker version $E\{\exp(\theta' A_n - \theta' C_n \theta/2)\} \leq 1$ for $\|\theta\| < \varepsilon$, and other variants given in Chap. 10 provide a general set of conditions from which we can derive exponential bounds and moment inequalities for self-normalized processes in dependent settings. A key tool is the *pseudo-maximization* method which involves Laplace's method for evaluating integrals of the form $\int_{\|\theta\|<\varepsilon} e^{\theta' A_n - \theta' C_n \theta/2} f(\theta) d\theta$. If the random function $\exp\{\theta' A_n - \theta' C_n \theta/2\}$ in (1.2) could be maximized over θ inside the expectation $E\{\exp(\theta' A_n - \theta' C_n \theta/2)\}$, taking the maximizing value $\theta = C_n^{-1} A_n$ would yield the expectation of the self-normalized variable $\exp\{A_n C_n^{-1} A_n/2\}$. Although this argument is not valid, integrating $\exp\{\theta' A_n - \theta' C_n \theta/2\}$ with respect to $f(\theta) d\theta$ and applying Laplace's method to evaluate the integral basically achieves the same effect as in the heuristic argument. This method is used to derive exponential and L_p-bounds for self-normalized processes in Chap. 12. The exponential bounds are used to derive laws of the iterated logarithm for self-normalized processes in Chap. 13.

Student's t-statistic $\sqrt{n}(\bar{X}_n - \mu)/s_n$ has also undergone far-reaching generalizations in the statistics literature during the past century. Its generalization is the *Studentized statistic* $(\hat{\theta}_n - \theta)/\widehat{\mathrm{se}}_n$, where θ is a functional $g(F)$ of the underlying distribution function F, $\hat{\theta}_n$ is usually chosen to be the corresponding functional $g(\hat{F}_n)$ of the empirical distribution, and $\widehat{\mathrm{se}}_n$ is a consistent estimator of the standard error of $\hat{\theta}_n$. Its multivariate generalization, which replaces $1/\widehat{\mathrm{se}}_n$ by $\hat{\Sigma}_n^{-1/2}$, where $\hat{\Sigma}_n$ is a consistent estimator of the covariance matrix of the vector $\hat{\theta}_n$ or its variant, is ubiquitous in statistical applications. Part III of the book, which is on statistical applications of self-normalized processes, begins with an overview in Chap. 15 of the distribution theory of the t-statistic and its multivariate extensions, for samples first from normal distributions and then from general distributions that may have infinite second moments. Chapter 15 also considers the asymptotic theory of general Studentized statistics in time series and control systems and relates this theory to that of self-normalized martingales. An alternative to inference based on asymptotic distributions of Studentized statistics is to make use of bootstrapping. Chapter 16 describes the role of self-normalization in deriving approximate pivots for the construction of bootstrap confidence intervals, whose accuracy and correctness are analyzed by Edgeworth and Cornish–Fisher expansions. Chapter 17 introduces generalized likelihood ratio statistics as another class of self-normalized statistics. It also relates the pseudo-maximization approach and the method of mixtures in Part II to the close connections between likelihood and Bayesian inference. Whereas the framework of Part I covers the classical setting of independent observations sampled from a population, that of Part II is applicable to time series models and stochastic dynamic systems, and examples are given in Chaps. 15, 17 and 18. Moreover, the probability theory in Parts I and II is related not only to samples of fixed size, but also to sequentially generated samples that are associated with asymptotically optimal stopping rules. Part III concludes with Chap. 18 which considers self-normalized processes in sequential analysis and the associated boundary crossing problems.

Part I
Independent Random Variables

Chapter 2
Classical Limit Theorems, Inequalities and Other Tools

This chapter summarizes some classical limit theorems, basic probability inequalities and other tools that are used in subsequent chapters. Throughout this book, all random variables are assumed to be defined on the same probability space $(\Omega, \mathscr{F}, P)$ unless otherwise specified.

2.1 Classical Limit Theorems

The law of large numbers, the central limit theorem and the law of the iterated logarithm form a trilogy of the asymptotic behavior of sums of independent random variables. They are closely related to moment conditions and deal with three modes of convergence of a sequence of random variables Y_n to a random variable Y. We say that Y_n converges to Y *in probability*, denoted by $Y_n \xrightarrow{P} Y$, if, for any $\varepsilon > 0$, $P(|Y_n - Y| > \varepsilon) \to 0$ as $n \to \infty$. We say that Y_n converges *almost surely* to Y (or Y_n converges to Y with probability 1), denoted by $Y_n \xrightarrow{a.s.} Y$, if $P(\lim_{n\to\infty} Y_n = Y) = 1$. Note that almost sure convergence is equivalent to $P(\max_{k \ge n} |Y_k - Y| > \varepsilon) \to 0$ as $n \to \infty$ for any given $\varepsilon > 0$. We say that Y_n converges *in distribution* (or *weakly*) to Y, and write $Y_n \xrightarrow{D} Y$ or $Y_n \Rightarrow Y$, if $P(Y_n \le x) \to P(Y \le x)$, at every continuity point of the cumulative distribution function of Y. If the cumulative distribution $P(Y \le x)$ is continuous, then $Y_n \xrightarrow{D} Y$ not only means $P(Y_n \le x) \to P(Y \le x)$ for every x, but also implies that the convergence is uniform in x, i.e.,

$$\sup_x |P(Y_n \le x) - P(Y \le x)| \to 0 \qquad \text{as } n \to \infty.$$

The three modes of convergence are related by

$$Y_n \xrightarrow{a.s.} Y \Longrightarrow Y_n \xrightarrow{P} Y \Longrightarrow Y_n \xrightarrow{D} Y.$$

V.H. de la Peña et al., *Self-Normalized Processes: Limit Theory and Statistical Applications*, Probability and its Applications,

The reverse relations are not true in general. However, $Y_n \xrightarrow{D} c$ is equivalent to $Y_n \xrightarrow{P} c$ when c is a constant. Another relationship is provided by Slutsky's theorem: If $Y_n \xrightarrow{D} Y$ and $\xi_n \xrightarrow{P} c$, then $Y_n + \xi_n \xrightarrow{D} Y + c$ and $\xi_n Y_n \xrightarrow{D} cY$.

2.1.1 The Weak Law, Strong Law and Law of the Iterated Logarithm

Let $X_1, X_2, \ldots$ be *independent and identically distributed* (i.i.d.) random variables and let $S_n = \sum_{i=1}^n X_i$. Then we have Kolmogorov's strong law of large numbers and Feller's weak law of large numbers.

Theorem 2.1. *$n^{-1}S_n \xrightarrow{a.s.} c < \infty$ if and only if $E(|X_1|) < \infty$, in which case $c = E(X_1)$.*

Theorem 2.2. *In order that there exist constants c_n such that $n^{-1}S_n - c_n \xrightarrow{P} 0$, it is necessary and sufficient that $\lim_{x\to\infty} xP(|X_1| \geq x) = 0$. In this case, $c_n = EX_1 I(|X_1| \leq n)$.*

The Marcinkiewicz–Zygmund law of large numbers gives the rate of convergence in Theorem 2.1.

Theorem 2.3. *Let $1 < p < 2$. If $E(|X_1|) < \infty$, then*

$$n^{1-1/p}\left(n^{-1}S_n - E(X_1)\right) \xrightarrow{a.s.} 0 \tag{2.1}$$

if and only if $E(|X_1|^p) < \infty$.

When $p = 2$, (2.1) is no longer valid. Instead, we have the Hartman–Wintner *law of the iterated logarithm* (LIL), the converse of which is established by Strassen (1966).

Theorem 2.4. *If $EX_1^2 < \infty$ and $EX_1 = \mu$, $Var(X_1) = \sigma^2$, then*

$$\limsup_{n\to\infty} \frac{S_n - n\mu}{\sqrt{2n\log\log n}} = \sigma \quad a.s.,$$

$$\liminf_{n\to\infty} \frac{S_n - n\mu}{\sqrt{2n\log\log n}} = -\sigma \quad a.s.,$$

$$\limsup_{n\to\infty} \frac{\max_{1\leq k\leq n}|S_k - k\mu|}{\sqrt{2n\log\log n}} = \sigma \quad a.s.$$

Conversely, if there exist finite constants a and τ such that

$$\limsup_{n\to\infty} \frac{S_n - na}{\sqrt{2n\log\log n}} = \tau \quad a.s.,$$

then $a = E(X_1)$ and $\tau^2 = Var(X_1)$.

The following is an important tool for proving Theorems 2.1, 2.3 and 2.4.

Lemma 2.5 (Borel–Cantelli Lemma).

(1) Let $A_1, A_2, \ldots$ be an arbitrary sequence of events on $(\Omega, \mathscr{F}, P)$. Then $\sum_{i=1}^{\infty} P(A_i) < \infty$ implies $P(A_n \text{ i.o.}) = 0$, where $\{A_n \text{ i.o.}\}$ denotes the event $\bigcap_{k\geq 1} \bigcup_{n\geq k} A_n$, i.e., A_n occurs infinitely often.
(2) Let $A_1, A_2, \ldots,$ be a sequence of independent events on $(\Omega, \mathscr{F}, P)$. Then $\sum_{i=1}^{\infty} P(A_i) = \infty$ implies $P(A_n \text{ i.o.}) = 1$.

The strong law of large numbers and LIL have also been shown to hold for independent but not necessarily identically distributed random variables $X_1, X_2, \ldots$.

Theorem 2.6.

(1) If $b_n \uparrow \infty$ and $\sum_{i=1}^{\infty} \mathrm{Var}(X_i)/b_i^2 < \infty$, then $(S_n - ES_n)/b_n \xrightarrow{a.s.} 0$.
(2) If $b_n \uparrow \infty$, $\sum_{i=1}^{\infty} P(|X_i| \geq b_i) < \infty$ and $\sum_{i=1}^{\infty} b_i^{-2} EX_i^2 I(|X_i| \leq b_i) < \infty$, then $(S_n - a_n)/b_n \xrightarrow{a.s.} 0$, where $a_n = \sum_{i=1}^{n} EX_i I(|X_i| \leq b_i)$.

The "if" part in Theorems 2.1 and 2.3 can be derived from Theorem 2.6, which can be proved by making use Kolmogorov's three-series theorem and the Kronecker lemma in the following.

Theorem 2.7 (Three-series Theorem). *The series $\sum_{i=1}^{\infty} X_i$ converges a.s. if and only if the three series*

$$\sum_{i=1}^{\infty} P(|X_i| \geq c), \quad \sum_{i=1}^{\infty} EX_i I(|X_i| \leq c), \quad \sum_{i=1}^{\infty} \mathrm{Var}\{X_i I(|X_i| \leq c)\}$$

converge for some $c > 0$.

Lemma 2.8 (Kronecker's Lemma). *If $\sum_{i=1}^{\infty} x_i$ converges and $b_n \uparrow \infty$, then $b_n^{-1} \sum_{i=1}^{n} b_i x_i \to 0$.*

We end this subsection with Kolmogorov's LIL for independent but not necessarily identically distributed random variables; see Chow and Teicher (1988, Sect. 10.2). *Assume that $EX_i = 0$ and $EX_i^2 < \infty$ and put $B_n^2 = \sum_{i=1}^{n} EX_i^2$. If $B_n \to \infty$ and $X_n = o(B_n(\log\log B_n)^{-1/2})$ a.s., then*

$$\limsup_{n\to\infty} \frac{S_n}{B_n\sqrt{2\log\log B_n}} = 1 \ \ a.s. \tag{2.2}$$

2.1.2 The Central Limit Theorem

For any sequence of random variables X_i with finite means, the sequence $X_i - E(X_i)$ has zero means and therefore we can assume, without loss of generality, that the mean of X_i is 0. For i.i.d. X_i, we have the classical central limit theorem (CLT).

Theorem 2.9. *If $X_1,\ldots,X_n$ are i.i.d. with $E(X_1)=0$ and $\mathrm{Var}(X_1)=\sigma^2<\infty$, then*

$$\frac{S_n}{\sqrt{n}\,\sigma} \xrightarrow{D} N(0,1).$$

The Berry–Esseen inequality provides the convergence rate in the CLT.

Theorem 2.10. *Let Φ denote the standard normal distribution function and $W_n = S_n/(\sqrt{n}\sigma)$. Then*

$$\begin{aligned}
&\sup_x |P(W_n \le x) - \Phi(x)| \qquad (2.3)\\
&\le 4.1\left\{\sigma^{-2}EX_1^2 I\left(|X_1|>\sqrt{n}\sigma\right) + n^{-1/2}\sigma^{-3}E|X_1|^3 I\left(|X_1|\le\sqrt{n}\sigma\right)\right\}.
\end{aligned}$$

In particular, if $E|X_1|^3<\infty$, then

$$\sup_x |P(W_n\le x)-\Phi(x)| \le \frac{0.79E|X_1|^3}{\sqrt{n}\sigma^3}. \qquad (2.4)$$

For general independent not necessarily identically distributed random variables, the CLT holds under the Lindeberg condition, under which a non-uniform Berry–Esseen inequality of the type in (2.3) still holds.

Theorem 2.11 (Lindberg–Feller CLT). *Let X_n be independent random variables with $E(X_i)=0$ and $E(X_i^2)<\infty$. Let $W_n=S_n/B_n$, where $B_n^2=\sum_{i=1}^n E(X_i^2)$. If the Lindberg condition*

$$B_n^{-2}\sum_{i=1}^n EX_i^2 I(|X_i|\ge\varepsilon B_n) \longrightarrow 0 \qquad \text{for all } \varepsilon>0 \qquad (2.5)$$

holds, then $W_n \xrightarrow{D} N(0,1)$. Conversely, if $\max_{1\le i\le n} EX_i^2 = o(B_n^2)$ and $W_n \xrightarrow{D} N(0,1)$, then the Lindberg condition (2.5) is satisfied.

Theorem 2.12. *With the same notations as in Theorem 2.11,*

$$\begin{aligned}
&\sup_x |P(W_n\le x)-\Phi(x)| \qquad (2.6)\\
&\le 4.1\left(B_n^{-2}\sum_{i=1}^n EX_i^2 I\{|X_i|>B_n\} + B_n^{-3}\sum_{i=1}^n E|X_i|^3 I\{|X_i|\le B_n\}\right)
\end{aligned}$$

and

$$\begin{aligned}
&|P(W_n\le x)-\Phi(x)| \qquad (2.7)\\
&\le C\left(\sum_{i=1}^n \frac{EX_i^2 I\{|X_i|>(1+|x|)B_n\}}{(1+|x|)^2B_n^2} + \sum_{i=1}^n \frac{E|X_i|^3 I\{|X_i|\le(1+|x|)B_n\}}{(1+|x|)^3B_n^3}\right),
\end{aligned}$$

where C is an absolute constant.

2.1.3 Cramér's Moderate Deviation Theorem

The Berry–Esseen inequality gives a bound on the absolute error in approximating the distribution of W_n by the standard normal distribution. The usefulness of the bound may be limited when $\Phi(x)$ is close to 0 or 1. Cramér's theory of moderate deviations provides the relative errors. Petrov (1975, pp. 219–228) gives a comprehensive treatment of the theory and introduces the *Cramér series*, which is a power series whose coefficients can be expressed in terms of the cumulants of the underlying distribution and which is used in part (a) of the following theorem.

Theorem 2.13.

(a) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E(X_1) = 0$ and $Ee^{t_0|X_1|} < \infty$ for some $t_0 > 0$. Then for $x \geq 0$ and $x = o(n^{1/2})$,

$$\frac{P(W_n \geq x)}{1-\Phi(x)} = \exp\left\{x^2 \lambda\left(\frac{x}{\sqrt{n}}\right)\right\}\left(1 + O\left(\frac{1+x}{\sqrt{n}}\right)\right), \tag{2.8}$$

where $\lambda(t)$ is the Cramér series.

(b) If $Ee^{t_0\sqrt{|X_1|}} < \infty$ for some $t_0 > 0$, then

$$\frac{P(W_n \geq x)}{1-\Phi(x)} \to 1 \qquad \text{as } n \to \infty \text{ uniformly in } x \in \left[0, o(n^{1/6})\right). \tag{2.9}$$

(c) The converse of (b) is also true; that is, if (2.9) holds, then $Ee^{t_0\sqrt{|X_1|}} < \infty$ for some $t_0 > 0$.

In parts (a) and (b) of Theorem 2.13, $P(W_n \geq x)/(1-\Phi(x))$ can clearly be replaced by $P(W_n \leq -x)/\Phi(-x)$. Moreover, similar results are also available for standardized sums S_n/B_n of independent but not necessarily identically distributed random variables with bounded moment generating functions in some neighborhood of the origin; see Petrov (1975). In Chap. 7, we establish Cramér-type moderate deviation results for *self-normalized* (rather than standardized) sums of independent random variables under much weaker conditions.

2.2 Exponential Inequalities for Sample Sums

2.2.1 Self-Normalized Sums

We begin by considering independent Rademacher random variables.

Theorem 2.14. *Assume that ε_i are independent and $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$. Then*

$$P\left(\frac{\sum_{i=1}^n a_i \varepsilon_i}{\left(\sum_{i=1}^n a_i^2\right)^{1/2}} \geq x\right) \leq e^{-x^2/2} \tag{2.10}$$

for $x > 0$ and real numbers $\{a_i\}$.

Proof. Without loss of generality, assume $\sum_{i=1}^n a_i^2 = 1$. Observe that

$$\frac{1}{2}(e^{-t} + e^t) \le e^{t^2/2}$$

for $t \in \mathbb{R}$. We have

$$\begin{aligned} P\left(\sum_{i=1}^n a_i\varepsilon_i \ge x\right) &\le e^{-x^2} E e^{x\sum_{i=1}^n a_i\varepsilon_i} \\ &= e^{-x^2} \prod_{i=1}^n \frac{1}{2}(e^{-a_i x} + e^{a_i x}) \\ &\le e^{-x^2} \prod_{i=1}^n e^{a_i^2 x^2/2} = e^{-x^2/2}. \end{aligned}$$

□

Let X_n be independent random variables and let $V_n^2 = \sum_{i=1}^n X_i^2$. If we further assume that X_i is symmetric, then X_i and $\varepsilon_i X_i$ have the same distribution, where $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables independent of $\{X_i\}$. Hence the self-normalized sum S_n/V_n has the same distribution as $(\sum_{i=1}^n X_i\varepsilon_i)/V_n$. Given $\{X_i, 1 \le i \le n\}$, applying (2.10) to $a_i = X_i$ yields the following.

Theorem 2.15. *If X_i is symmetric, then for $x > 0$,*

$$P(S_n \ge xV_n) \le e^{-x^2/2}. \tag{2.11}$$

The next result extends the above "sub-Gaussian" property of the self-normalized sum S_n/V_n to general (not necessarily symmetric) independent random variables.

Theorem 2.16. *Assume that there exist $b > 0$ and a such that*

$$P(S_n \ge a) \le 1/4 \quad \textit{and} \quad P(V_n^2 \ge b^2) \le 1/4. \tag{2.12}$$

Then for $x > 0$,

$$P\{S_n \ge x(a + b + V_n)\} \le 2e^{-x^2/2}. \tag{2.13}$$

In particular, if $E(X_i) = 0$ and $E(X_i^2) < \infty$, then

$$P\{|S_n| \ge x(4B_n + V_n)\} \le 4e^{-x^2/2} \qquad \textit{for } x > 0, \tag{2.14}$$

where $B_n = (\sum_{i=1}^n EX_i^2)^{1/2}$.

Proof. When $x \le 1$, (2.13) is trivial. When $x > 1$, let $\{Y_i, 1 \le i \le n\}$ be an independent copy of $\{X_i, 1 \le i \le n\}$. Then

$$\begin{aligned} P\left(\sum_{i=1}^n Y_i \le a, \sum_{i=1}^n Y_i^2 \le b^2\right) &\ge 1 - P\left(\sum_{i=1}^n Y_i > a\right) - P\left(\sum_{i=1}^n Y_i^2 > b^2\right) \\ &\ge 1 - 1/4 - 1/4 = 1/2. \end{aligned}$$

Noting that

$$\begin{aligned}
&\left\{S_n \geq x(a+b+V_n),\ \sum_{i=1}^n Y_i \leq a, \sum_{i=1}^n Y_i^2 \leq b^2\right\} \\
&\subset \left\{\sum_{i=1}^n (X_i - Y_i) \geq x\left(a+b+\left(\sum_{i=1}^n (X_i-Y_i)^2\right)^{1/2} - \left(\sum_{i=1}^n Y_i^2\right)^{1/2}\right) - a, \sum_{i=1}^n Y_i^2 \leq b^2\right\} \\
&\subset \left\{\sum_{i=1}^n (X_i - Y_i) \geq x\left(\sum_{i=1}^n (X_i-Y_i)^2\right)^{1/2}\right\}
\end{aligned}$$

and that $\{X_i - Y_i, 1 \leq i \leq n\}$ is a sequence of independent symmetric random variables, we have

$$\begin{aligned}
P\left(S_n \geq x(a+b+V_n)\right) &= \frac{P\left(S_n \geq x(a+b+V_n), \sum_{i=1}^n Y_i \leq a, \sum_{i=1}^n Y_i^2 \leq b^2\right)}{P\left(\sum_{i=1}^n Y_i \leq a, \sum_{i=1}^n Y_i^2 \leq b^2\right)} \\
&\leq 2P\left(\sum_{i=1}^n (X_i - Y_i) \geq x\left(\sum_{i=1}^n (X_i-Y_i)^2\right)^{1/2}\right) \\
&\leq 2e^{-x^2/2}
\end{aligned}$$

by (2.11). This proves (2.13), and (2.14) follows from (2.13) with $a = b = 2B_n$. $\square$

2.2.2 Tail Probabilities for Partial Sums

Let X_n be independent random variables and let $S_n = \sum_{i=1}^n X_i$. The following theorem gives the *Bennett–Hoeffding inequalities*.

Theorem 2.17. *Assume that $EX_i \leq 0$, $X_i \leq a$ $(a > 0)$ for each $1 \leq i \leq n$, and $\sum_{i=1}^n EX_i^2 \leq B_n^2$. Then*

$$Ee^{tS_n} \leq \exp\left(a^{-2}(e^{ta} - 1 - ta)B_n^2\right) \qquad \text{for } t > 0, \tag{2.15}$$

$$P(S_n \geq x) \leq \exp\left(-\frac{B_n^2}{a^2}\left\{\left(1+\frac{ax}{B_n^2}\right)\log\left(1+\frac{ax}{B_n^2}\right) - \frac{ax}{B_n^2}\right\}\right) \tag{2.16}$$

and

$$P(S_n \geq x) \leq \exp\left(-\frac{x^2}{2(B_n^2 + ax)}\right) \qquad \text{for } x > 0. \tag{2.17}$$

Proof. It is easy to see that $(e^s - 1 - s)/s^2$ is an increasing function of s. Therefore

$$e^{ts} \leq 1 + ts + (ts)^2(e^{ta} - 1 - ta)/(ta)^2 \tag{2.18}$$

for $s \le a$, and hence

$$Ee^{tS_n} = \prod_{i=1}^n Ee^{tX_i} \le \prod_{i=1}^n \left(1 + tEX_i + a^{-2}(e^{ta} - 1 - ta)EX_i^2\right)$$
$$\le \prod_{i=1}^n \left(1 + a^{-2}(e^{ta} - 1 - ta)EX_i^2\right) \le \exp\left(a^{-2}(e^{ta} - 1 - ta)B_n^2\right).$$

This proves (2.15). To prove (2.16), let $t = a^{-1}\log(1 + ax/B_n^2)$. Then, by (2.15),

$$\begin{aligned} P(S_n \ge x) &\le e^{-tx} Ee^{tS_n} \\ &\le \exp\left(-tx + a^{-2}(e^{ta} - 1 - ta)B_n^2\right) \\ &= \exp\left(-\frac{B_n^2}{a^2}\left\{\left(1 + \frac{ax}{B_n^2}\right)\log\left(1 + \frac{ax}{B_n^2}\right) - \frac{ax}{B_n^2}\right\}\right), \end{aligned}$$

proving (2.16). To prove (2.17), use (2.16) and

$$(1+s)\log(1+s) - s \ge \frac{s^2}{2(1+s)} \qquad \text{for } s > 0.$$

□

The inequality (2.17) is often called *Bernstein's inequality*. From the Taylor expansion of e^x, it follows that

$$e^x \le 1 + x + x^2/2 + |x|^3 e^x/6. \tag{2.19}$$

Let $\beta_n = \sum_{i=1}^n E|X_i|^3$. Using (2.19) instead of (2.18) in the above proof, we have

$$Ee^{tS_n} \le \exp\left(\frac{1}{2}t^2 B_n^2 + \frac{1}{6}t^3 \beta_n e^{ta}\right), \tag{2.20}$$

$$P(S_n \ge x) \le \exp\left(-tx + \frac{1}{2}t^2 B_n^2 + \frac{1}{6}t^3 \beta_n e^{ta}\right) \tag{2.21}$$

for all $t > 0$, and in particular

$$P(S_n \ge x) \le \exp\left(-\frac{x^2}{2B_n^2} + \frac{x^3}{6B_n^6}\beta_n e^{ax/B_n^2}\right). \tag{2.22}$$

When X_i is not bounded above, we can first truncate it and then apply Theorem 2.17 to prove the following inequality.

Theorem 2.18. *Assume that $EX_i \le 0$ for $1 \le i \le n$ and that $\sum_{i=1}^n EX_i^2 \le B_n^2$. Then*

$$P(S_n \ge x) \le P\left(\max_{1\le i\le n} X_i \ge b\right) + \exp\left(-\frac{B_n^2}{a^2}\left\{\left(1 + \frac{ax}{B_n^2}\right)\log\left(1 + \frac{ax}{B_n^2}\right) - \frac{ax}{B_n^2}\right\}\right)$$
$$+ \sum_{i=1}^n P(a < X_i < b) P(S_n - X_i > x - b) \tag{2.23}$$

for $x > 0$ and $b \geq a > 0$. In particular,

$$P(S_n \geq x) \leq P\left(\max_{1\leq i\leq n} X_i > \delta x\right) + \left(\frac{3B_n^2}{B_n^2 + \delta x^2}\right)^{1/\delta} \tag{2.24}$$

for $x > 0$ and $\delta > 0$.

Proof. Let $\bar{X}_i = X_i I(X_i \leq a)$ and $\bar{S}_n = \sum_{i=1}^n \bar{X}_i$. Then

$$\begin{aligned}
P(S_n \geq x) &\leq P\left(\max_{1\leq i\leq n} X_i \geq b\right) + P\left(S_n \geq x, \max_{1\leq i\leq n} X_i \leq a\right) \\
&\quad + P\left(S_n \geq x, \max_{1\leq i\leq n} X_i > a, \max_{1\leq i\leq n} X_i < b\right) \\
&\leq P\left(\max_{1\leq i\leq n} X_i \geq b\right) + P(\bar{S}_n \geq x) \\
&\quad + \textstyle\sum_{i=1}^n P(S_n \geq x, a < X_i < b) \qquad (2.25) \\
&\leq P\left(\max_{1\leq i\leq n} X_i \geq b\right) + P(\bar{S}_n \geq x) \\
&\quad + \textstyle\sum_{i=1}^n P(S_n - X_i \geq x - b, a < X_i < b) \\
&= P\left(\max_{1\leq i\leq n} X_i \geq b\right) + P(\bar{S}_n \geq x) \\
&\quad + \textstyle\sum_{i=1}^n P(a < X_i < b) P(S_n - X_i \geq x - b).
\end{aligned}$$

Applying (2.16) to $\bar{S}_n$ gives

$$P(\bar{S}_n \geq x) \leq \exp\left(-\frac{B_n^2}{a^2}\left[\left(1 + \frac{ax}{B_n^2}\right)\log\left(1 + \frac{ax}{B_n^2}\right) - \frac{ax}{B_n^2}\right]\right),$$

which together with (2.26) yields (2.23). From (2.23) with $a = b = \delta x$, (2.24) follows. □

The following two results are about nonnegative random variables.

Theorem 2.19. *Assume that $X_i \geq 0$ with $E(X_i^2) < \infty$. Let $\mu_n = \sum_{i=1}^n EX_i$ and $B_n^2 = \sum_{i=1}^n EX_i^2$. Then for $0 < x < \mu_n$,*

$$P(S_n \leq x) \leq \exp\left(-\frac{(\mu_n - x)^2}{2B_n^2}\right). \tag{2.26}$$

Proof. Note that $e^{-a} \leq 1 - a + a^2/2$ for $a \geq 0$. For any $t \geq 0$ and $x \leq \mu_n$, we have

$$\begin{aligned}
P(S_n \leq x) &\leq e^{tx} E e^{-tS_n} = e^{tx} \prod_{i=1}^n E e^{-tX_i} \\
&\leq e^{tx} \prod_{i=1}^n E(1 - tX_i + t^2 X_i^2/2) \\
&\leq \exp\left(-t(\mu_n - x) + t^2 B_n^2/2\right).
\end{aligned}$$

Letting $t = (\mu_n - x)/B_n^2$ yields (2.26). □

Theorem 2.20. *Assume that $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. Then for $x > 0$,*

$$P(S_n \geq x) \leq \left(\frac{\mu e}{x}\right)^x, \tag{2.27}$$

where $\mu = \sum_{i=1}^n p_i$.

Proof. Let $t > 0$. Then

$$\begin{aligned} P(S_n \geq x) &\leq e^{-tx} \prod_{i=1}^n E e^{tX_i} = e^{-tx} \prod_{i=1}^n \left(1 + p_i(e^t - 1)\right) \\ &\leq \exp\left(-tx + (e^t - 1)\textstyle\sum_{i=1}^n p_i\right) = \exp\left(-tx + (e^t - 1)\mu\right). \end{aligned}$$

Since the case $x \leq \mu$ is trivial, we assume that $x > \mu$. Then letting $t = \log(x/\mu)$ yields

$$\exp\left(-tx + (e^t - 1)\mu\right) = \exp\left(-x\log(x/\mu) + x - \mu\right) \leq (\mu e/x)^x.$$

□

We end this section with the Ottaviani maximal inequality.

Theorem 2.21. *Assume that there exists a such that* $\max_{1\leq k\leq n} P(S_k - S_n \geq a) \leq 1/2$. *Then*

$$P\left(\max_{1\leq k\leq n} S_k \geq x\right) \leq 2P(S_n \geq x - a). \tag{2.28}$$

In particular, if $E(X_i) = 0$ and $E(X_i^2) < \infty$, then

$$P\left(\max_{1\leq k\leq n} S_k \geq x\right) \leq 2P(S_n \geq x - \sqrt{2}B_n), \tag{2.29}$$

where $B_n = \sqrt{\sum_{i=1}^n E(X_i^2)}$.

Proof. Let $A_1 = \{S_1 \geq x\}$ and $A_k = \{S_k \geq x, \max_{1\leq i\leq k-1} S_i < x\}$. Then $\{\max_{1\leq k\leq n} S_k \geq x\} = \cup_{k=1}^n A_k$ and

$$\begin{aligned} P\left(\max_{1\leq k\leq n} S_k \geq x\right) &\leq P(S_n \geq x - a) + \sum_{k=1}^n P(A_k, S_n < x - a) \\ &\leq P(S_n \geq x - a) + \sum_{k=1}^n P(A_k, S_n - S_k < -a) \\ &= P(S_n \geq x - a) + \sum_{k=1}^n P(A_k)P(S_n - S_k < -a) \\ &\leq P(S_n \geq x - a) + (1/2)\sum_{k=1}^n P(A_k) \\ &= P(S_n \geq x - a) + (1/2)P\left(\max_{1\leq k\leq n} S_k \geq x\right), \end{aligned}$$

which gives (2.28). (2.29) follows from (2.28) with $a = \sqrt{2}B_n$. □

The proof of Kolmogorov's LIL (2.2) involves upper exponential bounds like those in Theorem 2.17 and the following lower exponential bound, whose proof is given in Chow and Teicher (1988, pp. 352–354) and uses the "conjugate method" that will be described in Sect. 3.1.

Theorem 2.22. *Assume that* $EX_i = 0$ *and* $|X_i| \le a_i$ *a.s. for* $1 \le i \le n$ *and that* $\sum_{i=1}^n EX_i^2 = B_n^2$. *Let* $c_n \ge c_0 > 0$ *be such that* $\lim_{n\to\infty} a_n c_n / B_n = 0$. *Then for every* $0 < \gamma < 1$, *there exists* $0 < \delta_\gamma < 1/2$ *such that for all large n,*

$$P\left\{S_n \ge (1-\gamma)^2 c_n B_n\right\} \ge \delta_\gamma \exp\left\{-(1-\gamma)(1-\gamma^2)c_n^2/2\right\}.$$

2.3 Characteristic Functions and Expansions Related to the CLT

Let Y be a random variable with distribution function F. The *characteristic function* of Y is defined by $\varphi(t) = Ee^{itY} = \int_{-\infty}^{\infty} e^{ity} dF(y)$ for $t \in \mathbb{R}$. In view of *Lévy's inversion formula*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)dt = P(a < Y < b) + \frac{1}{2}\{P(Y=a) + P(Y=b)\} \tag{2.30}$$

for $a < b$ (see Durrett, 2005, pp. 93–94), the characteristic function uniquely determines the distribution function. The characteristic function φ is continuous, with $\varphi(0) = 1$, $|\varphi(t)| \le 1$ for all $t \in \mathbb{R}$. There are three possibilities concerning solutions to the equation $|\varphi(t)| = 1$ (see Durrett, 2005, p. 129):

(a) $|\varphi(t)| < 1$ for all $t \neq 0$.
(b) $|\varphi(t)| = 1$ for all $t \in \mathbb{R}$. In this case, $\varphi(t) = e^{ita}$ and Y puts all its mass at a.
(c) $|\varphi(\tau)| = 1$ and $|\varphi(t)| < 1$ for $0 < t < \tau$. In this case $|\varphi|$ has period τ and there exists $b \in \mathbb{R}$ such that the support of Y is the lattice $\{b + 2\pi j/\tau : j = 0, \pm 1, \pm 2, \dots\}$, i.e., Y is *lattice with span* $2\pi/\tau$.

A random variable Y is called *non-lattice* if its support is not a lattice, which corresponds to case (a) above. It is said to be *strongly non-lattice* if it satisfies *Cramér's condition*

$$\limsup_{|t|\to\infty} |\varphi(t)| < 1. \tag{2.31}$$

Note that (2.31), which is only concerned with the asymptotic behavior of $|\varphi(t)|$ as $|t| \to \infty$, is stronger than (a) because it rules out (b) and (c).

If the characteristic function φ of Y is integrable, i.e., $\int_{-\infty}^{\infty} |\varphi(t)|dt < \infty$, then Y has a bounded continuous density function f with respect to Lebesgue measure and

$$f(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \varphi(t)dt. \tag{2.32}$$

This is the *Fourier inversion formula*; see Durrett (2005, p. 95). In this case, since $\varphi(t) = \int_{-\infty}^{\infty} e^{ity} f(y)dy$ and f is integrable,

$$\lim_{|t|\to\infty} \varphi(t) = 0 \tag{2.33}$$

by the Riemann–Lebesgue lemma; see Durrett (2005, p. 459). Hence, if Y has an integrable characteristic function, then Y satisfies Cramér's condition (2.31).

In the case of lattice distributions with support $\{b + hk : k = 0, \pm 1, \pm 2, \dots\}$, let $p_k = P(Y = b + hk)$. Then the characteristic function is a Fourier series $\varphi(t) = \sum_{k=-\infty}^{\infty} p_k e^{it(b+hk)}$, with

$$p_k = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} e^{-it(b+hk)} \varphi(t)dt, \tag{2.34}$$

noting that the span h corresponds to $2\pi/\tau$ (or $\tau = 2\pi/h$) in (b).

2.3.1 Continuity Theorem and Weak Convergence

Theorem 2.23. *Let φ_n be the characteristic function of Y_n.*

(a) If $\varphi_n(t)$ converges, as $n \to \infty$, to a limit $\varphi(t)$ for every t and if φ is continuous at 0, then φ is the characteristic function of a random variable Y and $Y_n \Rightarrow Y$.
(b) If $Y_n \Rightarrow Y$ and φ is the characteristic function of Y, then $\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$ for all $t \in \mathbb{R}$.

For independent random variables $X_1, \dots, X_n$, the characteristic function of the sum $S_n = \sum_{k=1}^{n} X_k$ is the product of their characteristic functions $\varphi_1, \dots, \varphi_n$. If X_i has mean 0 and variance σ_i^2, quadratic approximation of $\varphi_i(t)$ in a neighborhood of the origin by Taylor's theorem leads to the central limit theorem under the Lindeberg condition (2.5). When the X_k have infinite second moments, the limiting distribution of $(S_n - b_n)/a_n$, if it exists for suitably chosen centering and scaling constants, is an *infinitely divisible* distribution, which is characterized by the property that its characteristic function is the nth power of a characteristic function for every integer $n \geq 1$. Equivalently, Y is infinitely divisible if for every $n \geq 1$, $Y \stackrel{D}{=} X_{n1} + \cdots + X_{nn}$, where X_{ni} are i.i.d. random variables and $\stackrel{D}{=}$ denotes equality in distribution (i.e., both sides having the same distribution). Another equivalent characterization of infinite divisibility is the Lévy–Khintchine representation of the characteristic function φ of Y:

$$\varphi(t) = \exp\left\{ i\gamma t + \int_{-\infty}^{\infty} \left(e^{itu} - 1 - \frac{itu}{1+u^2} \right) \left(\frac{1+u^2}{u^2} \right) dG(u) \right\}, \tag{2.35}$$

where $\gamma \in \mathbb{R}$ and G is nondecreasing, left continuous with $G(-\infty) = 0$ and $G(\infty) < \infty$. Examples of infinitely divisible distributions include the normal,