

Springer Texts in Statistics

Jiming Jiang

# Large Sample Techniques for Statistics

*Second Edition*

 Springer

# Springer Texts in Statistics

## Series Editors

G. Allen, Rice University, Department of Statistics, Houston, TX, USA

R. De Veaux, Department of Mathematics and Statistics, Williams College,  
Williamstown, MA, USA

R. Nugent, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA,  
USA

*Springer Texts in Statistics (STS)* includes advanced textbooks from 3rd- to 4th-year undergraduate courses to 1st- to 2nd-year graduate courses. Exercise sets should be included. The series editors are currently Genevera I. Allen, Richard D. De Veaux, and Rebecca Nugent. Stephen Fienberg, George Casella, and Ingram Olkin were editors of the series for many years.

More information about this series at <https://link.springer.com/bookseries/417>

Jiming Jiang

# Large Sample Techniques for Statistics

Second Edition

 Springer

Jiming Jiang  
Department of Statistics  
University of California, Davis  
Davis, CA, USA

ISSN 1431-875X                      ISSN 2197-4136 (electronic)  
Springer Texts in Statistics  
ISBN 978-3-030-91694-7              ISBN 978-3-030-91695-4 (eBook)  
<https://doi.org/10.1007/978-3-030-91695-4>

1<sup>st</sup> edition: © Springer Science+Business Media, LLC 2010

2<sup>nd</sup> edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*For my parents, Huifen and Haoliang,  
and my sisters, Qiuming and Dongming,  
with love*

# Preface

A quote from the preface of the first edition: “Large-sample techniques provide solutions to many practical problems; they simplify our solutions to difficult, sometimes intractable problems; they justify our solutions; and they guide us to directions of improvements.”

A lot of changes have taken place in the world, including in the world of statistics, now often referred to as *data science*, since the publication of the first edition. The changes are exclusively driven by practical needs. Nowadays, it has become increasingly easier to collect data, not only in the amount but also in features of the data, such as high-dimensional and graphical data, frequently updated over the internet. In terms of the amount of data, once a luxury hope, large-sample has now become a customary, if not yet necessary, feature of the data. Thus, in a way, large-sample techniques are becoming increasingly important. Below are some examples.

1. *Subject-level inference.* Random effects models, or more generally, mixed effects models, are often used when there is insufficient data or information, at the subject level. Examples include longitudinal data analysis, in which data are collected from individuals over time, and small area estimation (see Chapter 13 of the first edition). Here, the individuals or small areas are what we call subjects. As data collection has become increasingly feasible, the once small or moderate sample sizes at the individual or small area level may no longer be small—they are going to “infinity” as well, using a large-sample term. The ability to collect data at the subject level has led to new scientific frontiers, such as precision medicine. The National Research Council of the United States in 2014 defined the latter as the “ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those disease they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not.” Another area, in modern economic studies, is family economics, which applies basic economic concepts to families or small firms. It is not surprising that statistical

inference with primary interest at the subject levels has become one of the hottest topics of data science.

2. *Functional data.* Traditional statistics dealt with numbers and vectors. Many forms of modern data are functional, such as graphical or imaging data. For example, the trajectory of the recovery of a patient after receiving a medical treatment is a function of time. In the traditional analysis of longitudinal data, the observations were assumed to be collected at a finite set of time points. But, if the patient is constantly monitored, the change over time is similar to a continuous curve, and there is a different curve for a different patient. Imaging data share similar features but at higher dimensions. A basic unit of imaging data is called a pixel. An image consists of a large, or huge, number of pixels. Thus, in a way, imaging data is different from a (continuous) trajectory curve in that the dimension of all the pixels combined is still finite. But this is a very high-dimensional form of data as in the next example.
3. *High-dimensional data.* There are many other types of high-dimensional data, other than the imaging data. For example, Genome-wide association study (GWAS), which typically refers to examination of associations between up to millions of genetic variants in the genome and certain traits of interest among unrelated individuals, has been very successful for detecting genetic variants that affect complex human traits/diseases in the past fifteen years. The genetic variants in the genome are located at the single-nucleotide polymorphisms, or SNPs. A typical GWAS data set may involve thousands, or tens of thousands of individuals, but the number of SNPs is much larger, ranging from hundreds of thousands to millions. In other words, the sample size,  $n$ , which corresponds to the number of individuals, is much smaller than the dimension of the unknown parameters,  $p$ , which may be regression coefficients associated with the SNPs. It has become a frequent feature of modern data science problems that the dimension of the parameter is larger, sometimes much larger, than the sample size.

Are the current large sample techniques ready for the new data science challenges? Yes or no. On the one hand, basic techniques, such as those introduced in the first six chapters of the first edition, are still fundamental for the missions of large-sample techniques, quoted at the beginning of this preface, for modern data science. Even for the special topics covered in Chapters 7–15 of the first edition, many of those techniques are still essential, although in some cases there is a need for further development. For example, the first example of subject-level inference mentioned above is closely related to the materials covered in Chapters 12 and 13 of the first edition. The functional data problems in the second example are closely related to the topics of Chapter 7 of the first edition, in particular.

On the other hand, there are useful large sample techniques for modern data science that were not covered in the first edition. One of these techniques is random matrix theory. A new chapter, Chapter 16, is added to cover random matrix theory and some of its applications in statistics. In particular, such topics are closely related to high-dimensional data, as mentioned above in the third example. The applications



include GWAS, estimation of large covariance matrices, high-dimensional linear models and time series.

Furthermore, there have been new developments on topics covered in the first edition, such as an open problem regarding consistency of the maximum likelihood estimator in generalized linear mixed models with crossed random effects. The open problem was solved two years after the first edition. These developments are included in the new edition.

Other changes include correction of typos found since the publication of the first edition. A number of students, whose names the author unfortunately cannot accurately recall, made contributions to those corrections.

In addition, the current edition has added more to the already large number of exercises in the first edition. The exercises are attached to each chapter and closely related to the materials covered, giving the readers plenty of opportunities to digest the materials and practice what they have learned. The new edition is mostly self-contained with the appendices providing relevant backgrounds in matrix algebra, measure theory, and mathematical statistics. A list of notation is also provided in the appendices for convenience.

The book is intended for a wide audience, ranging from senior undergraduate students to researchers with Ph.D. degrees. More specifically, Chapters 1–5 and parts of Chapters 10–15 are intended for senior undergraduate and M.S. students. For Ph.D. students and researchers, all chapters are suitable. A first course in mathematical statistics and a course in calculus are prerequisites. As it is unlikely that all 16 chapters will be covered in a single-semester or quarter-course, the following combinations of chapters are recommended for a single-semester course, depending on the focus of interest (for a single-quarter course some adjustment is necessary):

For a senior undergraduate or M.S.-level course on large sample techniques, Chapters 1–6.

For those interested in linear models, generalized linear models, mixed effects models, statistical genetics such as GWAS, and other related applications, Chapters 1–6, 8, 12, and parts of 16.

For those interested in time series, stochastic processes, and their applications, Chapters 1–6, 8–10, and parts of 16.

For those interested in semi-parametric, nonparametric statistics, functional data and their applications, Chapters 1–7 and 11.

For those interested in high-dimensional data, Chapters 1–6, 12 and 16.

For those interested in empirical Bayes methods, small-area estimation, subject-level inference and related fields, Chapters 1–6, 12, and 13.

For those interested in resampling methods, Chapters 10–7, 11, and 14.

For those interested in Monte Carlo methods and their applications in Bayesian inference, Chapters 1–6, 10, and 15.

For those interested in spatial statistics, Chapters 1–6, 9, and 10.

Thus, in particular, Chapters 1–6 are vital to any sequence recommended.

The author would like to restate his gratefulness to the colleagues, former supervisor and students, who have help with the first edition. In addition, the author

would like to thank Professor Debashis Paul for valuable suggestions regarding the random matrix theory covered in the newly added Chapter [16](#).

Davis, CA, USA  
September, 2021

Jiming Jiang

# Contents

<b>1</b>	<b>The <math>\epsilon</math>-<math>\delta</math> Arguments</b> .....	1
1.1	Introduction .....	1
1.2	Getting used to the $\epsilon$ - $\delta$ arguments .....	2
1.3	More examples .....	4
1.4	Case study: Consistency of MLE in the i.i.d. case .....	8
1.5	Some useful results .....	11
1.5.1	Infinite sequence .....	12
1.5.2	Infinite series .....	13
1.5.3	Topology .....	14
1.5.4	Continuity, differentiation, and integration .....	14
1.6	Exercises .....	17
<b>2</b>	<b>Modes of Convergence</b> .....	21
2.1	Introduction .....	21
2.2	Convergence in probability .....	22
2.3	Almost sure convergence .....	25
2.4	Convergence in distribution .....	28
2.5	$L^p$ convergence and related topics .....	33
2.6	Case study: $\chi^2$ -test .....	40
2.7	Summary and additional results .....	47
2.8	Exercises .....	49
<b>3</b>	<b>Big <math>O</math>, Small <math>o</math>, and the Unspecified <math>c</math></b> .....	55
3.1	Introduction .....	55
3.2	Big $O$ and small $o$ for sequences and functions .....	56
3.3	Big $O$ and small $o$ for vectors and matrices .....	59
3.4	Big $O$ and small $o$ for random quantities .....	62
3.5	The unspecified $c$ and other similar methods .....	67
3.6	Case study: The baseball problem .....	71
3.7	Case study: Likelihood ratio for a clustering problem .....	75
3.8	Exercises .....	81

<b>4</b>	<b>Asymptotic Expansions</b> .....	87
4.1	Introduction .....	87
4.2	Taylor expansion .....	89
4.3	Edgeworth expansion; method of formal derivation .....	96
4.4	Other related expansions .....	101
4.4.1	Fourier series expansion .....	101
4.4.2	Cornish–Fisher expansion .....	105
4.4.3	Two time series expansions .....	108
4.5	Some elementary expansions .....	110
4.6	Laplace approximation .....	114
4.7	Case study: Asymptotic distribution of the MLE .....	119
4.8	Case study: The Prasad–Rao method .....	123
4.9	Exercises .....	129
<b>5</b>	<b>Inequalities</b> .....	137
5.1	Introduction .....	137
5.2	Numerical inequalities .....	138
5.2.1	The convex function inequality .....	138
5.2.2	Hölder’s and related inequalities .....	141
5.2.3	Monotone functions and related inequalities .....	144
5.3	Matrix inequalities .....	149
5.3.1	Nonnegative definite matrices .....	149
5.3.2	Characteristics of matrices .....	152
5.4	Integral/moment inequalities .....	156
5.5	Probability inequalities .....	163
5.6	Case study: Some problems on existence of moments .....	171
5.7	Case study: A variance inequality .....	176
5.8	Exercises .....	180
<b>6</b>	<b>Sums of Independent Random Variables</b> .....	191
6.1	Introduction .....	191
6.2	The weak law of large numbers .....	192
6.3	The strong law of large numbers .....	197
6.4	The central limit theorem .....	201
6.5	The law of the iterated logarithm .....	207
6.6	Further results .....	212
6.6.1	Invariance principles in CLT and LIL .....	212
6.6.2	Large deviations .....	217
6.7	Case study: The least squares estimators .....	222
6.8	Exercises .....	227
<b>7</b>	<b>Empirical Processes</b> .....	235
7.1	Introduction .....	235
7.2	Glivenko–Cantelli theorem and statistical functionals .....	237
7.3	Weak convergence of empirical processes .....	240
7.4	LIL and strong approximation .....	243

7.5	Bounds and large deviations .....	245
7.6	Non-i.i.d. observations .....	248
7.7	Empirical processes indexed by functions .....	251
7.8	Case study: Estimation of ROC curve and ODC.....	253
7.9	Exercises .....	256
<b>8</b>	<b>Martingales</b> .....	<b>259</b>
8.1	Introduction.....	259
8.2	Examples and simple properties .....	261
8.3	Two important theorems of martingales.....	267
	8.3.1 The optional stopping theorem .....	267
	8.3.2 The martingale convergence theorem .....	270
8.4	Martingale laws of large numbers .....	273
	8.4.1 A weak law of large numbers .....	273
	8.4.2 Some strong laws of large numbers .....	274
8.5	A martingale central limit theorem and related topic.....	278
8.6	Convergence rate in SLLN and LIL .....	283
8.7	Invariance principles for martingales .....	286
8.8	Case study: CLTs for quadratic forms.....	289
8.9	Case study: Martingale approximation .....	295
8.10	Exercises .....	298
<b>9</b>	<b>Time and Spatial Series</b> .....	<b>305</b>
9.1	Introduction.....	305
9.2	Autocovariances and autocorrelations.....	309
9.3	The information criteria.....	312
9.4	ARMA model identification .....	316
9.5	Strong limit theorems for i.i.d. spatial series.....	321
9.6	Two-parameter martingale differences .....	323
9.7	Sample ACV and ACR for spatial series .....	327
9.8	Case study: Spatial AR models .....	330
9.9	Exercises .....	334
<b>10</b>	<b>Stochastic Processes</b> .....	<b>339</b>
10.1	Introduction.....	339
10.2	Markov chains .....	341
10.3	Poisson processes .....	348
10.4	Renewal theory .....	353
10.5	Brownian motion .....	356
10.6	Stochastic integrals and diffusions .....	363
10.7	Case study: GARCH models and financial SDE.....	369
10.8	Exercises .....	374
<b>11</b>	<b>Nonparametric Statistics</b> .....	<b>379</b>
11.1	Introduction.....	379
11.2	Some classical nonparametric tests.....	382
11.3	Asymptotic relative efficiency .....	386

11.4	Goodness-of-fit tests	392
11.5	$U$ -statistics	396
11.6	Density estimation	405
11.7	Exercises	411
<b>12</b>	<b>Mixed Effects Models</b>	417
12.1	Introduction	417
12.2	REML: Restricted maximum likelihood	421
12.3	Linear mixed model diagnostics	429
12.4	Inference about GLMM	439
12.5	Mixed model selection	451
12.6	Exercises	459
<b>13</b>	<b>Small-Area Estimation</b>	465
13.1	Introduction	465
13.2	Empirical best prediction with binary data	467
13.3	The Fay–Herriot model	476
13.4	Nonparametric small-area estimation	485
13.5	Model selection for small-area estimation	493
13.6	Exercises	502
<b>14</b>	<b>Jackknife and Bootstrap</b>	507
14.1	Introduction	507
14.2	The jackknife	510
14.3	Jackknifing the MSPE of EBP	517
14.4	The bootstrap	526
14.5	Bootstrapping time series	536
14.6	Bootstrapping mixed models	545
14.7	Exercises	555
<b>15</b>	<b>Markov-Chain Monte Carlo</b>	561
15.1	Introduction	561
15.2	The Gibbs sampler	564
15.3	The Metropolis–Hastings algorithm	570
15.4	Monte Carlo EM algorithm	575
15.5	Convergence rates of Gibbs samplers	580
15.6	Exercises	587
<b>16</b>	<b>Random Matrix Theory</b>	593
16.1	Introduction	593
16.2	Fundamental theorems of RMT	595
16.3	Large covariance matrices	606
16.4	High-dimensional linear models	611
16.5	Genome-wide association study	615
16.6	Application to time series	623
16.7	Exercises	628

- Appendix A** ..... 633
  - A.1 Matrix algebra ..... 633
    - A.1.1 Numbers associated with a matrix ..... 633
    - A.1.2 Inverse of a matrix ..... 634
    - A.1.3 Kronecker products ..... 635
    - A.1.4 Matrix differentiation ..... 636
    - A.1.5 Projection ..... 636
    - A.1.6 Decompositions of matrices and eigenvalues ..... 637
  - A.2 Measure and probability ..... 639
    - A.2.1 Measures ..... 639
    - A.2.2 Measurable functions ..... 641
    - A.2.3 Integration ..... 643
    - A.2.4 Distributions and random variables ..... 645
    - A.2.5 Conditional expectations ..... 648
    - A.2.6 Conditional distributions ..... 650
  - A.3 Some results in statistics ..... 651
    - A.3.1 The multivariate normal distribution ..... 651
    - A.3.2 Maximum likelihood ..... 653
    - A.3.3 Exponential family and generalized linear models ..... 655
    - A.3.4 Bayesian inference ..... 656
    - A.3.5 Stationary processes ..... 658
  - A.4 List of notation and abbreviations ..... 660
  
- References** ..... 667
  
- Index** ..... 679