Springer Texts in Statistics

Jiming Jiang

Large Sample Techniques for Statistics

Second Edition

Springer

Springer Texts in Statistics

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

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 Jiming Jiang
September, 2021

Contents