Preface

When I was a graduate student more than twenty five years ago, I was struggling to read many statistical research papers. This is particularly true at the time when I had passed my Ph.D. qualification examination. The goal of this book is to make it easier for Ph.D. students and new researchers to embark in their research area. During the past 30 years, statistics has become more an applied and more diversified science. In response to this trend, I have tried to cover as many different topics as possible. My main research interest focuses on likelihood-based inferences, which includes parametric likelihood, biased sampling likelihood, semiparametric likelihood, empirical likelihood and Godambe's estimating function theory.

This book is devoted to biased sampling problems (also called choice-based sampling in econometric parlance) and over-identified parameter estimation problems. When a proper randomization cannot be achieved, the observed sample will not be representative of the population of interest. This biased sampling problem appears frequently since in the real world, truly random sampling is not easily achievable or practically feasible. Biased sampling problems appear in many areas of research, including medicine, epidemiology and public health, social sciences and economics. As pointed out by Prof. James Heckman (1979), the 2000 Nobel Laureate in Economics, "Sample selection bias may arise in practice for two reasons. First, there may be self selection by the individuals or data units being investigated. Second, sample selection decisions by analysts or data processors operate in much the same fashion as self selection". This book would be of interest to those who work in the health, biological, social and physical sciences, as well as those who are interested in survey methodology and other areas of statistical science, among others.

Due to its convenience and cost effectiveness, one of the most efficient designs in health sciences research is the case-control design. Under this design, individuals (called cases) with the condition of interest (for example, cancer) are sampled. Their risk profiles for the condition are collected. Then some controls (do not satisfy the condition of interest, for example cancer free) are enrolled along with their risk profiles are also recorded. Since the numbers of cases and controls are fixed by

researchers, the ratio of cases and controls does not necessarily match the one in the entire target population. Consequently, this selection bias can lead to a result that is different from what we would have gotten if we had enrolled the entire population. In missing data problems, the likelihood function based on only those individuals with complete data is a biased version of the targeted one. In capture and recapture studies, each animal may be captured with a probability proportional to its size. In observational casual studies, the choice of treatment and control depends on the baseline covariates, which may lead to biased results if simply using the two sample comparison methods. In finite population problems the proportional to size design is very popular in survey, which is a special type of biased sampling. Meanwhile the length-biased sampling is one of the most naturally occurred types of biased sampling. Length-biased data are clearly encountered in applications of renewal processes, etiologic studies, genome-wide linkage studies, epidemiologic cohort studies, cancer prevention trials, and studies of labor economy. In observational studies, a prevalent cohort design that draws samples from individuals with a condition or disease at the time of enrollment is generally more efficient and practical. The recruited patients who have already experienced an initiating event are followed prospectively for the failure event (e.g. disease progression or death) or are right censored. Under this sampling design, individuals with longer survival times measured from the onset of the disease are more likely to be included in the cohort, whereas those with shorter survival times are unconsciously excluded. Length-biased sample thereby manifests in various data sets, because the "observed" time intervals from initiation to failure within the prevalent cohort tend to be longer than those arising from the underlying distribution of the general population. Finding appropriate adjustments for the potential selection bias in analyzing length-biased data or more general biased sampling problems has been a long-standing statistical problem. Although we use a prevalent cohort study in medical applications here to illustrate length-biased data, it is apparent that similar issues caused by biased sampling are common in many potential applications and sampling designs. For example, biased sampling problem occurs frequently in stereology and estimating dark matter distribution in astronomy. Stereology is the study of three-dimensional properties of objects or matter usually observed two-dimensionally.

It is worth mentioning that biased sampling problems may occur even if the sampling is unbiased. In fact if we model only the density ratios but leave the baseline density arbitrary in multiple sample problems, then we end up with a biased sampling problem since those populations other than the baseline one can be treated as a biased version of the baseline population, where the selection bias functions are the density ratios. Therefore methods developed in biased sampling problems can be borrowed to make robust inference instead of the full parametric approach in multiple sample problems.

When a model is defined through more estimating functions than the free parameters, it becomes an over-identified parameter problem. This problem occurs naturally if there exists auxiliary information. For example, in survey sampling, summarized information is available from published reports. Meta analysis is an

exciting area to combine similar studies to achieve a more precise analysis. An effective statistical method is to synthesize estimating functions. The advantage of using estimating function approach over the full parametric likelihood is that one does not need to specify the full parametric model. Therefore this approach is robust against possible model mis-specifications. Moreover the likelihood-based approach may be very cumbersome in some complex setup, such as in finance dependent data and time series data problems. In statistical and econometric literature there are a few efficient methods to combine over-identified estimating functions, among them, generalized method of moments (GMM), empirical likelihood method and estimating function theory are the most popular ones.

The importance of biased sampling problem and over-identified parameter problem has been well recognized in statistics and econometrics. Among many other important contributions, econometricians Heckman and McFadden (2000) were awarded the Nobel prizes in economics for their fundamental contribution to selection sampling problems. Heckman was cited "for the development of theory and methods for analyzing selective samples," and McFadden was cited "for his development of theory and methods for analyzing discrete choice". The econometrician Hansen (2013) won the same prize for his contribution on the ground break research on GMM. In the introduction of his award, it was cited for "….His seminal paper, 'Large Sample Properties of Generalized-Methods of Moments Estimators', (1982 Econometrica) importantly altered the landscape for how empirical research is done in finance and macroeconomics".

Breslow (2003), one of the leading bio-statisticians in the world, argued that statisticians and epidemiologists have made similar contributions to medicine with their work on case-control studies, analysis of incomplete data and casual inference. In spite of repeated nominations of such eminent figures, the Nobel Prize in physiology and medicine has been never awarded for work in biostatistics or epidemiology. Nevertheless, this indicates that the biased sampling and over-identified parameter (GMM) problems are fundamentally important and have wide applications in different research disciplines.

In addition to biased sampling and over-identified parameter problems, I will discuss some popular models and methods in econometrics, epidemiology, statistics, and biostatistics, including, among others, the Manski's (1975) maximum rank score, Han's (1986) maximum rank correlation estimation, Cosslett's (1983) maximum likelihood estimator of the binary choice model under monotonic constraints, Godambe's optimal estimating function theory, Owen's (1988) empirical likelihood, Kullback–Leibler likelihood and entropy family, semiparametric genetic mixture models, Kou and Ying's (1997) i.i.d. representation of a non-central hypergeometric distribution, casual inference and missing data, Vardi (1985)'s multiplicative censoring model, multi-sample Wicksell corpuscle problem, capture and recapture models, case and control problems with prevalent cases, and inference under monotonic function constraints using a combination of the EM algorithm and pool adjacent violation algorithm.

In this book I will give a brief overview of parametric likelihood inference and survival analysis, which makes easier for graduate students to understand the recent

developments of nonparametric or semiparametric methods and survival analysis based on truncated data, length-biased sampling data or backward time data in prevalent cohort studies. There are many good textbooks on parametric inference, among others, for example, Cox and Hinkley (1974), Lehmann and Casella (1998), Lehmann and Romano (2005), Caselle and Berger (2008) and Shao (2003). Excellent survival analysis books, include, among others, Cox and Oakes (1984), Kalbfleisch and Prentice (2002), Lawless (2002), Lancaster (1992), Andersen, Borgan, Gill and Keiding (1993) and Fleming and Harrington (1994). Since this book is aimed towards graduate students from different areas, the emphasis of this book is on statistical reasoning but not on the detailed mathematical derivations. I have tried to make the theories as assessable as possible. The detailed more advanced mathematical theories such as modern empirical process theory and information bound calculations can be found, among others, such as Bickel, Klassen, Ritov and Wellner (1993), Van der Vaart and Wellner (2003), Kosorok (2008a, b) and Tsiatis (2006). Since biased sampling problems have a natural connection with the sampling probability proportional to size problems in finite population, we recommend excellent survey sampling books by Cochran (1977), Sarndal, Swensson and Wretman (1991) and Thompson (1997) to readers. Other than those statistical books, some outstanding econometric reference books include, among others, Amemiya (1985), Gourieroux and Monfori (1990) and Lancaster (1990). Klugman, Panjer and Willmot's (2004) book provides in-depth coverage of modeling techniques used throughout many branches of actuarial science. Sham's (1998) "Statistics in Human Genetics" is an excellent statistical genetic reference book to new researchers in human disease genetics.

For young researchers, finally, I hope this book can play a role called "Pao zhuan yinyu" in Chinese, which means "cast a brick to attract jade", or, to offer a few commonplace remarks by way of introduction so that others may come up with valuable opinions. It would be the greatest encouragement to me if students can apply some methods discussed in this book to solve their own practical problems.

Rockville, USA Jing Qin

http://www.springer.com/978-981-10-4854-8

Biased Sampling, Over-identified Parameter Problems and Beyond

Qin, J.

2017, XVI, 624 p. 5 illus., 1 illus. in color., Hardcover
ISBN: 978-981-10-4854-8