Jin Zhang

Mathematical Statistics

（数理统计）

SCIENCE PRESS
Beijing

Mathematical Statistics

（数理统计）

Jin Zhang

SCIENCE PRESS
Beijing

Responsible Editor: Yulong Hao

Published by Science Press
16 Donghuangchenggen North Street
Beijing 100717, P.R.China

Printed in Chengdu

ISBN 978-7-03-037409-7

Dedicated to my wife Jikun Yi
and my daughter Yili Zhang

In the memory of my father Junmo Zhang
and my mother Zhenying Zhu

Preface

This book is intended as a textbook or a reference book for a one-semester graduate or senior undergraduate course in mathematical statistics. It is written for students majoring in statistics or related fields.

Although there are many excellent English textbooks on this subject, most of them contain lengthy explanations and examples, which are difficult for non-native English readers to understand. My teaching experience in China and Canada has inspired me to write a textbook with simple language and concise examples, reducing the language barrier for students and teachers from non-English-speaking countries.

This book grew from my lecture notes developed for teaching mathematical statistics at Yunnan University (China) and University of Manitoba (Canada). The contents and structure of the book are mainly taken from the classical textbook *Mathematical Statistics: Basic Ideas and Selected Topics* (Vol I, 2nd ed. Prentice Hall, 2002) by P. J. Bickel and K. A. Doksum, with reference to other standard textbooks, such as *Mathematical Statistics* (Chapman & Hall/CRC, 2000) by K. Knight, *Statistical Inference* (2nd ed. Duxbury Press, 2002) by G. Casella and R. L. Berger, and *Introduction to Mathematical Statistics* (6th ed. Prentice Hall, 2005) by R. V. Hogg, J. W. Mckean and A. T. Craig.

The mathematical background necessary for this book is linear algebra and advance calculus (but no measure theory). It is assumed that the reader is familiar with basic probability theory and statistical principle. The main objective of this book is to build theoretical statistics by providing the essential material, which is as mathematically rigorous as possible, to help students understand the statistical background, thinking and methodology on a deep level.

This book consists of six chapters, containing the essentials of mathematical statistics. The important statistical concepts and terminologies are italicized and indexed. At the end of each chapter, we provide students with a number of exercises selected from different sources, most notably the text by Bickel and Doksum.

Chapter 1 introduces statistical models and principles, including Bayesian models, the framework of decision theory, and some discussion about prediction, sufficiency and exponential families. Chapter 2 discusses the methods of parameter estimation in parametrical models, especially the algorithm of computing the MLEs (maximum likelihood estimates) and related EM (Expectation/Maximization) algorithm.

Chapter 3 covers the optimality theory in parameter estimation, including Bayes principle, minimax principle, unbiased estimation, Lehmann-Sheffé theorem, Rao-Blackwell theorem, and the information inequality. Chapter 4 presents the basic concepts and theories of hypothesis testing and confidence intervals (regions), focusing on the Neyman-Pearson lemma, likelihood ratio tests, as well as the duality between confidence regions and hypothesis tests.

Chapter 5 deals with the large sample theories, discussing consistency of estimation, asymptotic theories based on the delta method, and asymptotic normality and efficiency of the MLEs. Finally, Chapter 6 further discusses the large sample theories in the multiparameter case, with emphasis on the asymptotic behavior of the MLEs, large sample tests and confidence regions, and large sample tests for categorical data.

The book also provides the readers with Table of Common Statistical Distributions in Appendix A, which includes the commonly used discrete distributions, continuous distributions and multivariate distributions. For each listed distribution, the table provides the detailed information about its pdf/pmf, moments, moment generating function, and important notes on associated distributions. Indeed, the table itself has rich contents as good reference materials on distribution.

Like many other textbooks in mathematical statistics, the commonly used statistical tables for the standard normal, t, χ^2 and F distributions are attached in Appendix B, where the table values are computed by using statistical software **R**, available at the **R**'s official web site

http://www.r-project.org

In writing this book, I received great contribution from many of my students. I take this opportunity to thank those graduate students who took Mathematical Statistics from me and helped me in typewriting and proof-reading the manuscript. Among them are Jie Li, Xiaojie Yang, Tianxia Ai, Hua Li, Yunqi Zhang, Xiaozhun Zhuang and Menglin Li.

I would like to sincerely thank Professors P. J. Bickel and K. A. Doksum for writing an excellent textbook, from which I greatly benefited. I am also very grateful for all kinds of help from my teachers, colleagues, friends and students, especially Xueren Wang and Niansheng Tang, Yunnan University; Yuehua Wu, York University; Jianxin Pan, University of Manchester; Xuming He and Peter Song, University of Michigan (Ann Arbor); Michael Stephens and Richard Lockhart, Simon Fraser University; James Fu, Liqun Wang and Xikui Wang, University of Manitoba; Gemai Chen, University of Calgary; Jiahua Chen, University of British Columbia; Keming Yu, Brunel University.

In addition, I would like to acknowledge the financial support of Yunnan University and the Natural Science Foundation of China (NSFC) for publishing this book.

Last, but not least, I would like to sincerely thank my wife Jikun Yi and my

daughter Yili Zhang from the bottom of my heart for their patience, understanding, encouragement and steadfast support.

Jin Zhang
School of Mathematics and Statistics
Yunnan University
Kunming, China

Contents

1 Statistical Models and Principles **1**

1.1 Statistical Models 1

1.1.1 Data and Models 1

1.1.2 Parameters and Statistics 3

1.2 Bayesian Models 4

1.3 The Framework of Decision Theory 6

1.3.1 Components of the Decision Theory 6

1.3.2 Bayes and Minimax Criteria 7

1.4 Prediction 8

1.5 Sufficiency 11

1.6 Exponential Families 16

1.6.1 The One-Parameter Case 16

1.6.2 The Multiparameter Case 18

1.6.3 Properties of Exponential Families 20

1.6.4 Conjugate Families of Prior Distributions 21

1.7 Exercises 23

2 Methods of Parameter Estimation **31**

2.1 Essentials of Point Estimation 31

2.1.1 M-Estimation 31

2.1.2 The Substitution Principle 34

2.2 Least Squares and Maximum Likelihood Methods 36

2.2.1 Least Squares and Weighted Least Squares Estimation 36
2.2.2 Maximum Likelihood Estimation 39
2.3 The MLE in Exponential Families 43
2.4 Algorithmic Issues for Parameter Estimation 44
2.4.1 The Bisection Method 45
2.4.2 The Coordinate Ascent Method 45
2.4.3 The Newton-Raphson Algorithm 46
2.4.4 The EM Algorithm 47
2.5 Exercises 50

3 Measures of Performance and Optimality 55

3.1 Bayes Principle 55
3.2 Minimax Principle 59
3.3 Unbiased Estimation 63
3.4 The Information Inequality 69
3.4.1 The One-Parameter Case 69
3.4.2 The Multiparameter Case 73
3.5 Exercises 75

4 Hypothesis Tests and Confidence Regions 82

4.1 The Framework of Hypothesis Testing 82
4.2 The Neyman-Pearson Test 85
4.3 Uniformly Most Powerful Tests 88
4.4 Confidence Intervals and Regions 91
4.5 The Duality between Confidence Regions and Hypothesis Tests . . . 96
4.6 Uniformly Most Accurate Confidence Bounds 97
4.7 Bayesian Formulation of Credible Regions 100
4.8 Prediction Intervals 102
4.9 Likelihood Ratio Tests 103
4.9.1 Introduction 103
4.9.2 One-Sample Problem for a Normal Distribution 104

4.9.3 Two-Sample Problem with Equal Variance 106
4.9.4 Two-Sample Problem with Unequal Variances 107
4.9.5 Likelihood Ratio Tests for Bivariate Normal Distributions . . 109
4.10 Exercises 110

5 Asymptotic Theories 118

5.1 Introduction 118
5.2 Consistency 119
5.2.1 Consistency in Estimation 119
5.2.2 Consistency of M-Estimates 121
5.3 Asymptotics Based on the Delta Method 121
5.3.1 The Delta Method for Approximations of Moments 122
5.3.2 The Delta Method for Approximations of Distributions . . . 123
5.4 Asymptotic Theory in One Dimension 126
5.4.1 Asymptotic Normality of M-Estimates 127
5.4.2 Asymptotic Normality and Efficiency of MLEs 129
5.4.3 One-Sided Tests and Confidence Intervals Based on the MLE 130
5.5 Asymptotic Theory of the Posterior Distribution 131
5.6 Exercises 134

6 Asymptotics in the Multiparameter Case 142

6.1 Asymptotic Normality in k Dimensions 142
6.1.1 Asymptotic Normality of M-Estimates 142
6.1.2 Asymptotic Normality and Efficiency of MLEs 144
6.2 Large-Sample Tests and Confidence Regions 145
6.2.1 Asymptotic Distribution of the Likelihood-Ratio Test Statistic 145
6.2.2 Wald's and Rao's Large-Sample Tests and Confidence Regions 148
6.3 Large-Sample Tests for Categorical Data 150
6.3.1 Goodness-of-Fit Tests for Multinomial Models 150
6.3.2 Goodness-of-Fit Tests for Composite Multinomial Models . . 153
6.3.3 The χ^2 Tests for Contingency Tables 153

6.4 Exercises 157

Appendix A: Table of Common Distributions 162

Appendix B: Statistical Tables 174

Table 1. The Standard Normal Distribution 175

Table 2. Distribution of t 176

Table 3. Distribution of χ^2 177

Table 4. Distribution of F 178

References 184

Index 186

Chapter 1

Statistical Models and Principles

Statistics is concerned with collecting data, analyzing data and interpreting data. Our task is to extract information from the data, draw some conclusions, and interpret the results. We do not consider the problem of data collection in this book, but take the data as given and focus on the methods of data analysis: statistical inference and decision theory.

The first chapter introduces some fundamental concepts of mathematical statistics, including statistical models, Bayesian methods, the framework of decision theory, prediction, sufficiency, and exponential families, which are essential for the material in other chapters.

1.1 Statistical Models

A statistical model is a set of probability distributions on the sample space, which are proposed to generate the sampled data. It is often convenient to index the probability distributions of a statistical model by a parameter. Thus, a parametric statistical model is just a parametric family of probability distributions proposed to generate the data. In practice, different kinds of statistical models are used to analyze data, interpret data, and predict the future.

1.1.1 Data and Models

In mathematical statistics, the ultimate object of our endeavor is to analyze the data, which comes from most studies and experiments, scientific or industrial, large scale or small scale. Statisticians draw useful information from the sampled data,

using everything they know. The particular angle of mathematical statistics is to view data as the outcome of random experiments that we model mathematically.

Example 1.1.1. *Sampling Inspection.* Consider a population of N elements, for instance, a shipment of manufactured items, where the proportion of defective items, θ, is unknown. When a sample of size n is drawn from the population without replacement and inspected, the sample space consists of the numbers $0, 1, \cdots, n$ corresponding to all possible number of defective items found.

On this space, a random variable X can be defined by

$$X(k) = k, \ k = 0, 1, \cdots, n,$$

where k is number of defective items in the sample. Then, X has a *hypergeometric distribution*:

$$P(X = k) = \frac{\begin{pmatrix} N\theta \\ k \end{pmatrix} \begin{pmatrix} N - N\theta \\ n - k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

if $\max\{n - N(1-\theta), 0\} \leqslant k < \min\{N\theta, n\}$. The hypergeometric distribution model is denoted by $H(N\theta, N, n)$. □

Example 1.1.2. *One-Sample Models.* Let the sampled data be n independent measurements $x_1, x_2, \cdots, x_n$ of a physical constant μ, which are realizations of *independent and identically distributed* (i.i.d.) random variables $X_1, X_2, \cdots, X_n$ with common unknown distribution function F. Then, our model is

$$X_i = \mu + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$ is the vector of random error, satisfying the following assumptions:

(1) The distribution of $\boldsymbol{\varepsilon}$ is independent of μ.

(2) Random errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are mutually independent.

(3) Random errors $\varepsilon_1, \varepsilon_2 \cdots, \varepsilon_n$ are identically distributed.

(4) The common distribution of the random errors is normal with mean 0 and variance σ^2, which is unknown.

That is, random errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are i.i.d. from $N(0, \sigma^2)$. In other words, $X_1, X_2, \cdots, X_n$ are i.i.d. from $N(\mu, \sigma^2)$. Then the common distribution function of X_i's is $F(x) = \Phi(\frac{x - \mu}{\sigma})$, where $\Phi(x)$ is the distribution function of the standard normal distribution $N(0, 1)$. □

In general, suppose we have a random experiment and define a random variable X in the *sample space* $\mathcal{X}$, which is the set of all possible outcomes of the random

experiment. In the sample space $\mathcal{X}$, we observe the data $x_1, x_2, \cdots, x_n$, which are randomly drawn from $\mathcal{X}$ and can be thought of as outcomes or realizations of random variables $X_1, X_2, \cdots, X_n$ from $\mathcal{X}$ with some probability distribution P. In mathematical statistics, such $X_1, X_2, \cdots, X_n$ are known as a *random sample* from P if they are i.i.d. random variables with common probability distribution P.

Suppose that the joint distribution of the sample $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ is unknown but belongs to some family of probability distributions called a *statistical model*. It is often convenient to index the distributions of a statistical model by some *parameter* θ, which is a real number or vector and represents the unknown part of the model. Then we can write our statistical model as

$$\boldsymbol{X} = (X_1, X_2, \cdots, X_n) \sim P_\theta, \ \theta \in \Theta,$$

where "$\sim$" stands for "is distributed as", P_θ denotes the probability distribution or measure of the model indexed by parameter θ, and Θ is the *parameter space*, the set of all possible values for the parameter θ.

In particular, the statistical model in Example 1.1.1 is

$$X \sim H(N\theta, N, n), \ \theta \in \Theta,$$

where the sample space is $\mathcal{X} = \{0, 1, \cdots, n\}$, X is the number of defective items in the sample, the parameter θ is the proportion of defective items in the population, and the parameter space is $\Theta = \{0, \frac{1}{N}, \frac{2}{N}, \cdots, \frac{N-1}{N}, 1\}$.

The statistical model in Example 1.1.2 is

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2), \ \theta \in \Theta,$$

where the sample space is $\mathcal{X} = (-\infty, \infty)$, $X_1, X_2, \cdots, X_n$ are i.i.d. measurements of μ, $\theta = (\mu, \sigma^2)$ is the vector of parameters, and the parameter space is $\Theta = (-\infty, \infty) \times (0, \infty)$.

1.1.2 Parameters and Statistics

In statistics, the data from the experiment and the parameter of a statistical model have different roles. Actually, the data is observed by the experimenter, while the true parameter in the statistical model is unknown to the experimenter. Thus, the main goal of statistical analysis is to use the information from observed data to make inference about the unknown parameter of the model.

When we write $\boldsymbol{X} \sim P_\theta$, $\theta \in \Theta$, our sampled data $\boldsymbol{X}$ comes from a population whose distribution P_θ is indexed by a real-valued parameter (or parameter vector) θ, which captures important features of the population.

Usually we assume that Θ is a subset of Euclidean space, and the population distribution P_θ is completely specified when θ is known. Such a model is called a

parametric model. For example, we have a parametric model

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

in Example 1.1.2, where $\theta = (\mu, \sigma^2)$. However, this model becomes a *semiparametric model* if we drop the normality assumption about the underlying distribution.

To rule out the possibility that the parametrization is not one-to-one, we often require that the parametrization is *identifiable*, that is, $\theta_1 \neq \theta_2$ implies that $P_{\theta_1} \neq P_{\theta_2}$. Otherwise, the parametrization is called *unidentifiable.*

Under the assumption that $\boldsymbol{X} \sim P_\theta$, $\theta \in \Theta$, expectations calculated will be written as E_θ, cumulative distribution functions (cdf) will be denoted by $F(\cdot, \theta)$, and *probability density functions* (pdf) or *probability mass functions* (pmf) by $f(\cdot, \theta)$.

For any parametric model, it is convenient to consider either

(1) All of the probability distribution P_θ are continuous with pdf $f(x, \theta)$, where $\int f(x, \theta)dx = 1$ for all $\theta \in \Theta$;

(2) All of P_θ are discrete with pmf $f(x, \theta)$, the sample space $\{x_1, x_2, x_3, \cdots\}$ is independent of θ, and $\sum_{i=1}^{\infty} f(x_i, \theta) = 1$ for all $\theta \in \Theta$.

Such a model is known as a *regular model.* In Example 1.1.2 with assumptions (1)-(4), $\theta = (\mu, \sigma^2)$, $\Theta = R \times R^+$, and $f(x, \theta) = \prod_{i=1}^{n} \frac{1}{\sigma}\varphi(\frac{x_i - \mu}{\sigma})$, where φ is the density function of the standard normal distribution.

Statistics are functions defined on the sample space, which can be real-valued or vector-valued. Formally, a *statistic* is defined as a function of the sample: $T = T(\boldsymbol{X})$ that does not depends on any unknown parameter.

The aim of using a statistic is to summarize the information in the sampled data $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$. For example, to estimate the population mean μ and variance σ^2, we can use the simplest statistics

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

as their natural estimates, where $\overline{X}$ and s^2 are called the *sample mean* and *sample variance.*

1.2 Bayesian Models

There are different philosophies of statistical inference, which can be classified into two major schools: the *Frequentist* school and *Bayesian* school, representing the

classical (traditional) school and the modern (contemporary) school respectively. The Frequentist method is the most commonly used in practice, but it is, by no means, superior or inferior to the Bayesian method.

In statistics, Frequentists draw objective information from the sampled data to make inference on the unknown parameters of statistical models, while Bayesians use the sampled data to update subjective belief about uncertain parameters in the models.

The main criticism of Bayesian statistics concerns its subjectivity. Most critics of Bayesian methods focus on the difficulty of choosing the prior distribution. Nevertheless, the Bayesian approach provides a unified theory for statistical inference and decision making. As a matter of fact, every Frequentist result can be obtained using Bayesian methods by selecting a proper prior distribution.

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ be the sampled data. Consider a parametric model:

$$\boldsymbol{X} \sim P_\theta, \ \theta \in \Theta.$$

In the Frequentist approach, the parameter θ is considered to be unknown but fixed. Then the data $\boldsymbol{X}$ is the only information to be used to estimate the parameter θ or make inference about θ.

In the Bayesian approach, however, the parameter $\boldsymbol{\theta}$ is considered as a random variable or vector, which has a *prior distribution* with density or mass function $\pi(\theta)$, reflecting an experimenter's subject belief or information about the true parameter θ before the experiment. The prior is known as a *improper prior* if

$$\int \pi(\theta) d\theta = \infty \quad \text{or} \quad \sum \pi(\theta) = \infty.$$

The basic idea of Bayesian methods is to use the distribution of the sample to update the prior distribution.

In a Bayesian model, P_θ is considered as the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{\theta} = \theta$, so its pdf or pmf $f(\boldsymbol{x}, \theta)$ is denoted as $f(\boldsymbol{x}|\theta)$. Then, the joint pdf or pmf for $\boldsymbol{X}$ and $\boldsymbol{\theta}$ is

$$f(\boldsymbol{x}, \theta) = f(\boldsymbol{x}|\theta)\pi(\theta),$$

and the pdf or pmf of the *posterior distribution* is given by

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int f(\boldsymbol{x}|t)\pi(t)dt} \quad \text{if } \theta \text{ is continuous} \\ &= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\sum_t f(\boldsymbol{x}|t)\pi(t)} \quad \text{if } \theta \text{ is discrete,} \end{aligned}$$

which actually is the conditional pdf or pmf of $\boldsymbol{\theta}$ given $\boldsymbol{X} = \boldsymbol{x}$.

In Bayesian statistics, all the information about θ is contained in the posterior distribution, which combines the objective information on θ from the sampled data

and the subjective information from the prior distribution. Therefore, any statistical inference about θ should be decided by the posterior distribution.

The framework of the Bayesian decision theory will be briefly discussed in the next section. For more description and discussion about the Bayesian models, refer to Savage (1972) and Berger (1985).

1.3 The Framework of Decision Theory

In this section, we briefly introduce and discuss the basic concepts of *decision theory*. Generally speaking, the framework of decision theory consists of action space, loss function, decision rule and the risk function, as well as the Bayes and minimax criteria.

Decision theory provides a unified theory for statistical inference and decision making, enabling us to think clearly about estimation, testing, and confidence regions in a unified way.

1.3.1 Components of the Decision Theory

The decision theoretic foundation of statistics includes four elements: action space, loss function, decision rule, and risk function. We now introduce the notions of these four components of decision theory.

Action Space: $\mathcal{A}$

An *action space* $\mathcal{A}$ consists of actions (decisions or claims) that we can make based on the sampled data $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$. There are different types of action spaces for different situations, such as estimation of parameters and tests of hypotheses. Here are simple illustrations for the cases of estimation and testing.

Estimation. $\mathcal{A} = \{\text{all } \widehat{q(\theta)}\}$ for estimating $q(\theta)$, some function of θ, where $\widehat{q(\theta)}$ denotes any estimate of $q(\theta)$. For example, to estimate θ (the proportion of defective items) in Example 1.1.1, we can take $\mathcal{A} = \{0, \frac{1}{n}, \frac{2}{n}, \cdots, 1\}$, To estimate the population mean μ in Example 1.1.2, let $\mathcal{A} = R$.

Testing. $\mathcal{A} = \{0, 1\}$ with 0 or 1 corresponding to two actions: accepting or rejecting null hypothesis $H_0 : \theta \in \Theta_0$.

Loss function: $l(\theta, a)$

A *loss function* is used to describe the loss when we take an action a based on the observed sample $\boldsymbol{X} = \boldsymbol{x}$. It is defined as a function $l(\theta, a)$: $\Theta \times \mathcal{A} \to R^+$. The most commonly used lost functions are quadratic loss for estimation and 0-1 loss for testing, which are described as below.