# 复杂数据的bootstrap统计推断及其应用

徐礼文 著

科学出版社

# 复杂数据的 bootstrap 统计推断及其应用

徐礼文 著

科学出版社
北京

## 内 容 简 介

本书的主要内容是作者及其合作者在复杂数据模型这一领域近些年的研究成果, 以及相关的最新进展. 全书共 6 章. 第 1 章简要介绍几类复杂数据模型和 bootstrap 等预备知识和相关研究问题. 第 2~6 章, 系统讨论各种复杂数据统计推断中的 bootstrap 基本理论、方法及其应用, 包括 Behrens-Fisher 问题、异方差回归模型、异方差 ANOVA 和 MANOVA 模型、混合效应模型及高维数据分析中的 bootstrap 统计推断.

本书可供数理统计理论与方法研究的专家及数学、工程、经济、金融等领域的科研人员和工作者使用, 也可供数理统计专业的教师、高年级本科生及研究生作教材或参考书.

**图书在版编目(CIP)数据**

复杂数据的 bootstrap 统计推断及其应用/徐礼文著. —北京: 科学出版社, 2016. 7

ISBN 978-7-03-049523-5

Ⅰ. ①复… Ⅱ. ①徐… Ⅲ. ①数理统计 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2016) 第 187894 号

责任编辑: 李 欣 / 责任校对: 邹慧卿
责任印制: 张 伟 / 封面设计: 陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

北京教图印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2016 年 9 月第 一 版 开本: 720 × 1000 B5
2016 年 9 月第一次印刷 印张: 11 3/4
字数: 225 000

**定价: 69.00 元**

(如有印装质量问题, 我社负责调换)

# 前　言

复杂数据与复杂性科学相伴而生, 大量产生于许多科学研究和实践中, 包括常见的异方差数据、重复观测数据、空间分层数据、面板数据或纵向数据和高维海量数据等. 复杂数据的分析、建模和统计推断包含了当代几乎所有的统计研究分支, 并广泛应用在生物医学、经济金融和信息互联网技术中, 成为数据科学研究的重要组成部分. 现在, 由于数据的复杂性, 统计分析已无法离开计算机技术, 并为之提供学科发展动力. 特别是在如今这样一个信息爆炸性增长的大数据时代, 能够借助计算机技术, 以有效且富于统计意义的方式来处理海量复杂数据信息成为流行趋势. Bootstrap 方法是一种统计味道浓厚, 兼具计算机密集使用的强有力的统计技术. Bootstrap 方法的使用会让我们对统计学有更为深刻的理解, 而 bootstrap 方法的实施又离不开计算机技术的强力支撑. 本书在介绍复杂数据模型和 bootstrap 的基本理论与方法的基础上, 论述作者及其合作者近些年在异方差模型、重复观测数据模型、生长曲线模型、面板数据模型等 bootstrap 推断方面的研究工作, 以及其他一些与之紧密相关的最新研究进展. 因为论题涉及面较广且为作者的知识和书的篇幅所限, 只能着重从作者曾涉足而相对比较了解的领域来论述.

全书共 6 章, 第 1 章通过实例引进和比较各类复杂数据模型, 介绍模型的产生背景和本书重点应用到的 bootstrap 和广义推断等方面的知识. 第 2 章分别讨论了 Behrens-Fisher 问题中异方差下两个正态总体均值的相等性检验、均值差的置信区间和多个正态总体共同均值的参数 bootstrap 统计推断等问题. 第 3 章讨论了异方差情形下回归模型系数的相等性检验和共同回归系数的 PB 置信域. 第 4 章系统地研究了异方差下 ANOVA 中各种模型中的参数 bootstrap 检验和置信区间的构造方法和优良性. 第 5 章则将第 4 章的问题拓展到 MANOVA 模型, 并给出了 MANOVA 中检验的各种不变性理论. 最后一章转入混合效应模型和高维数据分析中的假设检验、置信区间和预测等的 bootstrap 方法研究.

本书的写作得到了国家自然科学基金项目 (项目编号: 11171002)、北京市自然科学基金项目 (项目编号: 1112008)、北京市属高等学校高层次人才引进与培养计划项目 (项目编号: CIT&TCD201404002) 和北方工业大学长城学者后备人才培养计划项目的资助, 作者愿借此机会表示诚挚的谢意!

限于作者水平, 书中不妥之处在所难免, 恳请读者不吝赐教.

徐礼文

2015 年 9 月于北京

# 符 号 表

| | |
|---|---|
| $A \geqslant 0$ | $A$ 为对称半正定方阵 |
| $A > 0$ | $A$ 为对称正定方阵 |
| $A^-$ | 矩阵 $A$ 的任一广义逆 |
| $A^+$ | 矩阵 $A$ 的 Moore-Penrose 广义逆 |
| $\mathrm{rank}(A)$ | 矩阵 $A$ 的秩 |
| $\mid A \mid$ | 矩阵 $A$ 的行列式 |
| $\parallel A \parallel$ | 矩阵 $A$ 的范数 |
| $\mathrm{tr}(A)$ | 方阵 $A$ 的迹 |
| $\mathscr{R}(A)$ | 矩阵 $A$ 的列向量张成的子空间 |
| $\mathscr{N}(A)$ | 矩阵 $A$ 的零空间 |
| $P_A$ | 向 $\mathscr{R}(A)$ 的正交投影变换阵 |
| $N_A$ | $N_A = I - P_A$ |
| $M^{\mathrm{T}}$ | 矩阵 $M$ 的转置矩阵 |
| $1_n = (1, \cdots, 1)^{\mathrm{T}}$ | 分量皆为 1 的 $n$ 维列向量 |
| $A \otimes B$ | $A$ 与 $B$ 的 Kronecker 乘积 |
| $E(X)$ | 随机变量或向量 $X$ 的均值 |
| $\mathrm{Var}(X)$ | 随机变量 $X$ 的方差 |
| $\mathrm{Cov}(X, Y)$ | 随机变量或向量 $X, Y$ 的协方差 |
| $u \sim N_p(\mu, \Sigma)$ | 均值为 $\mu$, 协方差阵为 $\Sigma$ 的 $p$ 维正态向量 |

# 目　　录

前言
符号表
第 1 章　引言 ····· 1
1.1　复杂数据及模型 ····· 1
1.1.1　Behrens-Fisher 问题 ····· 1
1.1.2　异方差回归模型 ····· 2
1.1.3　异方差的方差分析模型 ····· 2
1.1.4　生长曲线模型 ····· 2
1.1.5　Panel 数据模型 ····· 2
1.1.6　高维数据模型 ····· 4
1.2　复杂数据模型的相关研究进展 ····· 5
1.2.1　统计推断模式演化 ····· 5
1.2.2　分布推断方法的发展 ····· 6
1.3　Bootstrap 统计推断 ····· 8
1.3.1　Bootstrap 方法简介 ····· 8
1.3.2　Bootstrap $p$ 值检验 ····· 10
1.3.3　Bootstrap 置信区间 ····· 10
1.3.4　Bootstrap 光滑方法 ····· 11
1.4　广义推断 ····· 12
1.4.1　广义 $p$ 值 ····· 12
1.4.2　广义置信区间 ····· 12
第 2 章　Behrens-Fisher 问题的 bootstrap 解 ····· 14
2.1　引言 ····· 14
2.2　Behrens-Fisher 问题的参数 bootstrap 检验 ····· 14
2.2.1　均值相等性检验 ····· 15
2.2.2　模拟研究 ····· 18
2.3　两个正态总体均值差的 PB 区间估计 ····· 21
2.4　多个正态总体共同均值的参数 bootstrap 推断 ····· 23
2.4.1　引言 ····· 23

2.4.2 共同均值的推断 ····· 24
2.4.3 随机模拟研究 ····· 27
2.4.4 结论 ····· 32
**第 3 章 异方差回归模型中的 bootstrap 推断** ····· 33
3.1 引言 ····· 33
3.2 比较异方差回归模型的 PB 检验 ····· 33
3.3 异方差回归模型共同系数的 PB 置信域 ····· 35
**第 4 章 方差分析模型中 bootstrap 推断** ····· 38
4.1 单向方差分析模型 ····· 38
4.1.1 引言 ····· 38
4.1.2 PB 检验和 ADF 检验方法 ····· 39
4.1.3 数值结果 ····· 42
4.1.4 定理的证明 ····· 45
4.2 两向方差分析模型 (无交互效应) ····· 47
4.2.1 引言 ····· 47
4.2.2 固定效应模型检验 ····· 47
4.2.3 第一类错误概率和势函数性质 ····· 52
4.2.4 混合效应模型的检验 ····· 54
4.3 两向方差分析模型 (可能存在交互效应) ····· 55
4.3.1 引言 ····· 55
4.3.2 交互效应的检验 ····· 55
4.3.3 主效应的检验 ····· 60
4.3.4 数值结果 ····· 62
4.4 两因子套分类模型 ····· 65
4.4.1 引言 ····· 65
4.4.2 检验方法 ····· 65
4.4.3 因子 $A$ 的效应检验 ····· 69
4.4.4 模拟研究 ····· 72
4.4.5 两因子套设计模型随机套效应检验 ····· 74
4.4.6 实例分析 ····· 74
4.5 三因子套分类模型 ····· 75
4.5.1 引言 ····· 75
4.5.2 三因子套设计中固定效应的检验 ····· 76

4.5.3 因子 $A$ 和 $B$ 的效应检验 ····· 80
4.5.4 模拟研究 ····· 84
4.5.5 三因子套设计中随机套效应检验 ····· 87
4.5.6 一个实例 ····· 87
4.5.7 讨论 ····· 88
**第 5 章 多元方差分析模型中 bootstrap 推断** ····· 89
5.1 单向 MANOVA ····· 89
5.1.1 模型及预备知识 ····· 90
5.1.2 PB 检验 ····· 91
5.2 两向 MANOVA(无交互效应) ····· 92
5.2.1 引言 ····· 92
5.2.2 固定效应模型检验 ····· 93
5.2.3 数值结果 ····· 98
5.2.4 多元混合效应模型的检验 ····· 102
5.3 两向 MANOVA(可能存在交互效应) ····· 103
5.3.1 引言 ····· 103
5.3.2 检验方法 ····· 104
5.3.3 定理的证明 ····· 110
5.3.4 数值结果 ····· 115
5.4 多元套分类模型 ····· 120
5.4.1 引言 ····· 120
5.4.2 被嵌套效应的检验 ····· 121
5.4.3 嵌套效应的检验 ····· 126
5.4.4 Monte Carlo 研究 ····· 127
**第 6 章 混合效应模型和高维数据的 bootstrap 推断** ····· 131
6.1 引言 ····· 131
6.2 简单生长曲线模型中 bootstrap 推断 ····· 132
6.2.1 引言 ····· 132
6.2.2 固定效应和方差分量的两种推断 ····· 133
6.2.3 覆盖率和势函数的计算算法 ····· 139
6.2.4 数值结果 ····· 140
6.2.5 实例分析 ····· 143
6.3 Panel 数据模型中 bootstrap 推断 ····· 145

6.3.1 引言 ····· 145
6.3.2 单向误差分量回归模型的推断 ····· 145
6.3.3 覆盖率和势函数的算法 ····· 153
6.3.4 Monte Carlo 模拟研究 ····· 153
6.3.5 实际数据例子 ····· 156
6.3.6 两向误差分量回归模型 ····· 158
6.4 线性混合效应模型中 EBLUP 分布的 PB 近似 ····· 160
6.5 高维数据分析中的 PB 检验 ····· 162
6.5.1 资本资产定价模型 ····· 162
6.5.2 有效性假设 ····· 163
6.5.3 参数估计 ····· 163
6.5.4 PB 检验方法 ····· 164
6.6 Bootstrap 光滑与模型选择 ····· 164
6.6.1 引言 ····· 164
6.6.2 非参数 bootstrap 光滑 ····· 165
6.6.3 基于模型选择的模型平均 ····· 167
**参考文献** ····· 168

# 第1章 引 言

本章将简要介绍复杂数据分析常用的模型和面临的统计推断问题, 并给出一些例子, 使读者对问题的背景有所了解, 并对后面引进的观点、统计概念和理论方法有更深入理解. 另外, 我们在本章还将提供以后常用到的相关概念和预备知识.

## 1.1 复杂数据及模型

假设有 $n$ 个样本 $\{(x_i, y_i)\}_{i=1}^n$, 这里 $x_i = (x_{i1}, \cdots, x_{ip})$ 是 $p$ 维的预测变量或称作自变量, $y_i \in \mathbb{R}$ 是对应的响应变量. 经典的简单数据线性模型形式为

$$y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad \mathrm{Cov}(\varepsilon) = \sigma^2 I, \tag{1.1.1}$$

其中 $y = (y_1, \cdots, y_n)^{\mathrm{T}}$ 为 $n \times 1$ 的响应变量观测向量, $X$ 是第 $i$ 行为 $x_i$ 的 $n \times p$ 矩阵, $\beta$ 为 $p \times 1$ 的固定未知参数向量, $\varepsilon$ 是随机误差向量, $I$ 为 $n \times n$ 的单位阵. 从某种意义上说, 数据若不满足模型 (1.1.1) 中的某些假设条件则代表不同类型的复杂数据. 如 $\mathrm{Cov}(\varepsilon) = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2)$ 表示异方差数据; $\beta$ 中含有随机分量则可以建模分层数据、重复观测数据或面板数据等; 而 $n < p$ 则代表当今流行的高维数据问题. 下面我们简要介绍本书将要研究的几类复杂数据模型.

### 1.1.1 Behrens-Fisher 问题

最简单的一类复杂数据是带有异方差的两个总体比较问题的观测数据. 假设 $y_{i1}, \cdots, y_{in_i}$ 是来自正态总体 $N(\mu_i, \sigma_i^2)$ 的简单随机样本, $i = 1, 2$. 我们常感兴趣的问题是去检验假设 $H_0: \ \mu_1 = \mu_2$ 或寻找 $\mu_1 - \mu_2$ 的置信区间. 当总体方差任意且未知时, 文献中称该问题为 Behrens-Fisher 问题. 虽然此问题形式简单, 但至今没有一个标准解法, 已有许多学者提出各种解法. 为了以后的方便, 我们可以列出该数据的模型

$$\begin{aligned} y_1 &= 1_{n_1}\mu_1 + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_1^2 I_{n_1}), \\ y_2 &= 1_{n_2}\mu_2 + \varepsilon_2, \quad \varepsilon_2 \sim N(0, \sigma_2^2 I_{n_2}). \end{aligned} \tag{1.1.2}$$

模型 (1.1.2) 至少可以向两类复杂数据模型拓展, 一类是异方差回归模型, 另一类则是异方差的方差分析模型.

### 1.1.2　异方差回归模型

在诸如稳定性研究、刑侦学和变点分析等应用领域中, 许多学者考虑了正态线性模型回归系数的相等性检验问题. 若做模型的误差方差齐性假设, 该问题存在精确的 $F$ 检验. 但是, 在许多应用中齐性假设不可能成立或者难以验证. 一个自然的选择是下面的异方差回归模型:

$$\begin{aligned} y_1 &= X_1\beta_1 + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_1^2 I_{n_1}), \\ y_2 &= X_2\beta_2 + \varepsilon_2, \quad \varepsilon_2 \sim N(0, \sigma_2^2 I_{n_2}). \end{aligned} \tag{1.1.3}$$

这里 $X_i$ 是 $n_i \times p$ 的矩阵, $\beta_i$ 为 $p \times 1$ 的固定回归系数向量, $\varepsilon_i$ 是随机误差向量, $i = 1, 2$. 对于该模型, 我们感兴趣的问题是去检验假设 $H_0: \beta_1 = \beta_2$, 参见 Weerahandi(1987). 模型 (1.1.3) 还可以进一步推广到多个异方差回归模型和多元异方差的回归模型.

### 1.1.3　异方差的方差分析模型

异方差的方差分析是两个总体比较的 Behrens-Fisher 问题向多个总体比较的推广, 如异方差的单向方差分析模型、异方差的两向方差分析模型等. 常见的异方差单向方差分析模型为

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_i^2), \quad i = 1, \cdots, k; j = 1, \cdots, n_i, \tag{1.1.4}$$

这里 $\mu$ 是总平均值, $\alpha_i$ 为第 $i$ 个小总体的效应, 且 $\sum_{i=1}^{k} \alpha_i = 0$, $\varepsilon_{ij}$ 是随机误差. 对于该模型, 我们感兴趣的问题是去检验假设 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.

### 1.1.4　生长曲线模型

生长曲线 (growth-curve) 模型广泛应用于重复观测数据建模研究. 生物学家欲研究白鼠的某个特征随时间变化情况, 随机选用 $n$ 只小白鼠做试验. 设第 $i$ 只小白鼠的 $T$ 次观测值为 $y_{i1}, \cdots, y_{iT}$, $\tau_i$ 是和第 $i$ 个小白鼠相联系的个体随机效应, $X_t^{\mathrm{T}}$ 是 $t$ 时刻协变量的观测值向量, 例如, 对于 $p$ 阶多项式生长曲线模型, $X_t^{\mathrm{T}} = (1, t, \cdots, t^p), i = 1, \cdots, n$. 令 $\beta$ 是未知的系数向量, $\varepsilon_{it}$ 是误差项, 则

$$y_{it} = X_t^{\mathrm{T}}\beta + \tau_i + \varepsilon_{it}, \quad i = 1, \cdots, N;\ t = 1, \cdots, T. \tag{1.1.5}$$

这就是所谓的理论生长曲线, 其中 $\tau_i \sim N(0,\ \sigma_\tau^2)$, $\varepsilon_i \sim N(0,\ \sigma_\varepsilon^2)$. 生物学家的目的是估计 $\beta_0, \beta_1, \cdots, \beta_{p-1}$, 以得到经验生长曲线.

### 1.1.5　Panel 数据模型

这类模型被广泛用于计量经济学中 (Baltagi, 2013). 假设对 $N$ 个个体 (如个

人、家庭、公司、城市或国家) 进行了 $T$ 个时刻的观测, 观测数据可表示为

$$y_{it} = x_{it}^{\mathrm{T}}\beta + \xi_i + \varepsilon_{it}, \quad i = 1, \cdots, N;\ t = 1, \cdots, T, \tag{1.1.6}$$

其中 $y_{it}$ 表示第 $i$ 个个体第 $t$ 个时刻的某项经济指标, $x_{it}$ 是 $p \times 1$ 的已知向量, 它刻画了第 $i$ 个个体在第 $t$ 个时刻的一些自身特征, $\xi_i$ 是第 $i$ 个个体的个体效应, $\varepsilon_{it}$ 是随机误差项.

若我们的目的是研究整个市场的运行规律, 而不是关心这特定的 $N$ 个个体, 这 $N$ 个个体只不过是从总体中抽取的随机样本, 这时个体效应就是随机的. 我们假定所有 $\xi_i$ 和 $\varepsilon_{it}$ 互相独立, 且 $\xi_i \sim N(0,\ \sigma_\xi^2)$, $\varepsilon_i \sim N(0,\ \sigma_\varepsilon^2)$. 记

$$\begin{aligned} &y = (y_{11}, \cdots, y_{1T}, y_{21}, \cdots, y_{NT})^{\mathrm{T}}, \quad X = (x_{11}, \cdots, x_{1T}, x_{21}, \cdots, x_{NT})^{\mathrm{T}}, \\ &\xi = (\xi_1, \cdots, \xi_N)^{\mathrm{T}}, \quad \varepsilon = (\varepsilon_{11}, \cdots, \varepsilon_{1T}, \varepsilon_{21}, \cdots, \varepsilon_{NT})^{\mathrm{T}}, \quad U_1 = I_N \otimes 1_T, \end{aligned}$$

则模型 (1.1.6) 可以表示成

$$y = X\beta + U_1\xi + \varepsilon.$$

由上面的假设可知

$$\mathrm{Cov}(y) = \sigma_\xi^2 U_1 U_1^{\mathrm{T}} + \sigma_\varepsilon^2 I_{NT},$$

这里 $\sigma_\xi^2$ 和 $\sigma_\varepsilon^2$ 即为方差分量. 模型 (1.1.6) 有时也称为纵向数据 (longitudinal data) 模型, 常用于生物医药统计的研究领域中.

在上述问题中, 若把时间效应也考虑进来, 则模型 (1.1.6) 可以改写为

$$y_{it} = x_{it}^{\mathrm{T}}\beta + \xi_i + \eta_t + \varepsilon_{it}, \quad i = 1, \cdots, N;\ t = 1, \cdots, T,$$

其中 $\eta_t$ 表示与第 $t$ 个时刻相联系的时间效应, 若把其看成随机的, 并且假设 $\mathrm{Var}(\eta_t) = \sigma_\eta^2$, $\eta_t$ 与所有的 $\xi_i$ 和 $\varepsilon_{it}$ 相互独立, 记 $U_2 = 1_N \otimes I_T$, $\eta = (\eta_1, \cdots, \eta_T)^{\mathrm{T}}$, 则可得如下模型

$$y = X\beta + U_1\xi + U_2\eta + \varepsilon.$$

此时, 观测向量的协方差阵为

$$\mathrm{Cov}(y) = \sigma_\xi^2 U_1 U_1^{\mathrm{T}} + \sigma_\eta^2 U_2 U_2^{\mathrm{T}} + \sigma_\varepsilon^2 I_{NT},$$

其中 $\sigma_\xi^2, \sigma_\eta^2$ 和 $\sigma_\varepsilon^2$ 为方差分量.

上述的生长曲线模型和带有随机效应的 Panel 数据模型都是混合效应模型的特例, 其一般形式为

$$y = X\beta + Z\xi + \varepsilon, \tag{1.1.7}$$

这里 $y$ 为 $n\times 1$ 的观测向量, $\beta$ 为 $p\times 1$ 的非随机参数向量, 称为固定效应, $X$ 为对应于固定效应的设计阵; $\xi$ 为 $q\times 1$ 的随机向量, 称为随机效应, $Z$ 为对应于随机效应的设计阵; $\varepsilon$ 是随机误差向量. 一般假设 $E(\xi)=0$, $E(\varepsilon)=0$, $\xi$ 和 $\varepsilon$ 互不相关且

$$\mathrm{Cov}(\xi)=G,\quad \mathrm{Cov}(\varepsilon)=R.$$

于是 $\mathrm{Cov}(y)=\Sigma=ZGZ^{\mathrm{T}}+R$, 这里 $G$ 和 $R$ 分别为已知或未知的非负定和正定阵, 在它们未知时, 它们可以依赖于一个未知参数向量 $\theta$. 即此时模型 (1.1.7) 的随机部分 $Z\xi+\varepsilon$ 可以分解为 $Z\xi+\varepsilon=U_1\xi_1+U_2\xi_2+\cdots+U_k\xi_k$, 则得到一般的方差分量模型

$$y=X\beta+U_1\xi_1+U_2\xi_2+\cdots+U_k\xi_k, \tag{1.1.8}$$

其中 $\xi_i$ 为 $q_i\times 1$ 的随机效应向量, $U_i$ 为 $n\times q_i$ 的已知设计阵, 我们常常假设

$$E(\xi_i)=0,\quad \mathrm{Cov}(\xi_i)=\sigma_i^2 I_{q_i},\quad \mathrm{Cov}(\xi_i,\xi_j)=0,\quad i\neq j.$$

于是

$$E(y)=X\beta,\quad \mathrm{Cov}(y)=\sum_{i=1}^{k}\sigma_i^2U_iU_i^{\mathrm{T}},$$

$\sigma_i^2$ 称为方差分量, 最后一个随机效应向量 $\xi_k$ 是通常的随机误差向量 $\varepsilon$, 而 $U_k=I_n$. 模型 (1.1.8) 包含多类具有广泛应用背景的线性模型, 如 Panel 数据模型、单向 (两向、多向) 分类混合模型、套分类混合模型, 参见 Searle(1987), Rao 和 Kleffe(1988), Searle, Casella 和 McCulloch(1992), Diggle, Liang 和 Zeger(1994), Khuri 和 Mathew(1998), Verbeke 和 Molenberghs(2000) 以及 Baitagi(2013) 等.

### 1.1.6　高维数据模型

高维数据分析与建模是目前统计领域研究的热点之一 (Hastie, Tibshirani, Wainwright, 2015). 计算机技术的快速发展为人们存储数据带来了极大的便利, 所搜集数据的维数也呈几何级数的速度增长, 经常远远大于样本量的个数. 海量的数据为我们提供了更多的信息, 但与此同时, 也为如何进行数据分析提炼有效的信息带来了极大的挑战. 对于多元统计分析而言, 高维问题一般指如下两种情形: 一种是变量个数 $p$ 较大而样本量 $n$ 相对较小, 另一种是变量个数不大但是样本个数 $n$ 较多, 例如一项全国调查牵涉到大量的调查对象, 而观测指标个数相对较少. 面板数据高维问题较一般的多元高维问题更为复杂, 因为面板数据至少包括两个维度: 时间和横截面. 在实际应用中, 不同个体在不同时间进行观测时可以获得多个指标值. 可以用 $p$ 表示指标个数, $T$ 表示观测期长度, $N$ 表示个体个数. 统计学中所提到的高维问题, 通常是指个体数 $N$、时期长度 $T$ 或指标个数 $p$ 这三个变量中的一个或多个可以趋向于无穷, 描述该类问题的统计模型自然称为高维数据模型.

## 1.2 复杂数据模型的相关研究进展

针对复杂数据, 由于相应的模型假设条件放宽, 给实际的统计建模和统计推断带来不同程度的困难, 这些问题也推动复杂数据分析研究在许多方面取得新的进展, 获得了丰富多彩的研究成果. 下面仅对上述几类模型中的研究作简要介绍.

### 1.2.1 统计推断模式演化

随着处理数据的复杂程度的增加, 人们不断拓展经典的推断方式和看待参数的角度. 设随机变量 $X$ 是我们感兴趣总体的某个指标, 其分布函数记为 $F(x;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, $p$ 是参数维数, $\theta$ 是 $p$ 维参数, $\Theta$ 是参数空间. 若 $(x_1, \cdots, x_n)$ 是来自总体 $F(y;\theta)$ 的容量为 $n$ 的样本, 我们基于这些样本该如何推断参数? 杨振海 (2014) 指出, 基于看待参数的观点不同, 可以将文献中的推断模式概括为三种模式: 经典频率学派、Bayes 学派和信仰推断.

经典频率学派将参数看做固定常数. 样本 $x = (x_1, \cdots, x_n)$ 包含了参数 $\theta$ 的信息, 样本的分布函数是由参数 $\theta$ 确定的. 不过, 样本 $x$ 包含的信息无法精确确定参数 $\theta$ 的真实值, 但可以给出它的估计值, 如点估计和区间估计. 另外, 也可以根据区间估计, 在一定显著性水平下做出是否接受某参数 $\theta$ 等于某些值的决策, 即构造出合适的假设检验. 这些内容也是最常见的统计学方法.

Bayes 统计的基本观点如下. 设 $X \sim f(x;\theta)$, 其中 $\theta$ 是参数, 由于不同的 $\theta$ 值, $f(x;\theta)$ 对应着不同的密度函数, 从 Bayes 的观点看, 它是在给定 $\theta$ 值后的一个条件密度, 因此, 把 $f(x;\theta)$ 记为 $f(x \mid \theta)$ 更恰当些, 而 $f(x \mid \theta)$ 中关于 $\theta$ 的信息即为总体信息. 当 $\theta$ 给定后, 从总体 $f(x \mid \theta)$ 中随机地抽取一组样本 $X = (X_1, \cdots, X_n)$, 该样本中含有大量关于 $\theta$ 的信息, 这就是样本信息. 于是, 样本分布

$$f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

综合了总体和样本的信息.

我们对参数 $\theta$ 已经积累了一些资料, 经过加工整理后, 可以获得一些关于 $\theta$ 的有用信息, 这种信息称为 $\theta$ 的先验信息, 由于 $\theta$ 不是定值, 从 Bayes 观点看, 把 $\theta$ 当做一个随机变量. 关于 $\theta$ 的分布可以从先验信息中归纳出来, 故其分布称为先验分布, 用 $\pi(\theta)$ 表示其密度.

综合前面几点, 我们可以给出样本 $X = (X_1, \cdots, X_n)$ 和 $\theta$ 的联合密度

$$f(x, \theta) = f(x \mid \theta)\pi(\theta). \tag{1.2.1}$$

接下来的目的是, 利用联合分布, 即三方面的信息来推断 $\theta$. 当有了样本之后, 我们可以得到 $\theta$ 的条件分布:

$$\pi(\theta \mid x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}, \tag{1.2.2}$$

我们称之为 $\theta$ 的后验密度, 而 $f(x) = \int f(x \mid \theta)\pi(\theta)\mathrm{d}\theta$ 为样本 $X = (X_1, \cdots, X_n)$ 的边缘分布或样本的无条件分布.

可见, $\theta$ 的后验分布归纳了 $\theta$ 的先验信息、总体中和样本提供的信息. 于是, $\theta$ 的推断就应从后验分布入手. 常用的 Bayes 估计方法有:

取后验分布 $\pi(\theta \mid x)$ 的最大值点, 即众数作为 $\theta$ 估计, 称之为众数型 Bayes 估计.

取后验分布 $\pi(\theta \mid x)$ 的中位数作为 $\theta$ 估计, 称之为中位数型 Bayes 估计.

取后验分布 $\pi(\theta \mid x)$ 的期望作为 $\theta$ 估计, 称之为期望型 Bayes 估计.

除了占统治地位的频率学派和影响较大的 Bayes 学派, 近年来信仰推断也逐渐活跃, 并逐渐发展成现在的随机估计方法.

信仰推断是 Fisher 在极大似然估计思想的基础上提出来的. 极大似然估计方法是在有了样本 $x$ 之后, 用最大可能产生该样本的总体对应的参数值 $\hat{\theta}$ 去估计 $\theta$. Fisher(1930) 发展了这种思想, 认为观测到样本 $x$ 之后, 除了知道参数 $\theta$ 的可能取值外, 还确定出 $\theta$ 取这些值的可能性大小. 即此时又把参数 $\theta$ 看成了随机变量, 和经典统计产生了不一致, 也是大家广为争论的焦点. 但是, 换一种观点看待经典统计, 即将信仰推断归为分布推断, 则和 Bayes 推断一样, 都是用分布函数推断参数, 简称分布推断 (杨振海, 2014). 这也是现在部分学者提出置信分布 (confidence distribution, CD) 的基础.

### 1.2.2　分布推断方法的发展

最近, Xie, Singh 和 Strawderman(2011) 针对简单的参数模型, 提出了组合独立研究的框架, 并在此基础上只针对单个感兴趣参数, 发展出一个稳健的 Meta 分析方法, 其能有效地降低奇异研究 (outlying studies) 的影响. 他们提供组合方法时使用的工具就是置信分布. 置信分布指的是感兴趣参数空间上一个依赖于样本的分布函数, 作为感兴趣参数的分布估计, 能够提供感兴趣参数的点估计、所有水平的假设检验和置信区间. Singh, Xie 和 Strawderman(2005) 还给出了渐近 CD(asympotic confidence distribution, aCD) 的概念, 如非常有用的 bootstrap 分布就是一种 aCD. 事实上, CD 已经有了很长的历史 (Fisher, 1973; Efron, 1993). 但是, 最近的发展已经重新定义 CD 作为一个纯粹的频率学派的概念, 用来回答一个简单但重要的问题: 我们是否也能以类似 Bayes 后验分布推断的方式, 在频率推断中使用一个分布函数 (分布估计) 去估计感兴趣参数. 答案是肯定的.

历史上, CD 曾被理解成信仰分布 (fiducial distribution), 没有在频率推断框架

下得到完全发展. 最近, CD 又重新受到高度关注, 几项最新研究已经显示出其作为一种有效推断工具所蕴含的潜能, 参见 Schweder 和 Hjort(2002), Singh, Xie 和 Strawderman(2005), Xie, Singh 和 Strawderman(2011) 等. 他们的工作表明, CD 是参数信息的载体, 推断参数的依据和基础. CD 和信仰推断、Bayes 分析都是用分布函数推断参数. 另外, 因为在给定样本后, CD 就是感兴趣参数空间上一个完全已知的分布函数, 所以我们可以构造一个定义在样本空间和感兴趣参数空间乘积空间上的随机变量, 其在给定样本后的条件分布函数就是 CD. 该随机变量被称作待估参数的随机化估计, 简称随机估计, 如常见的 bootstrap 估计 (Xie, Singh, 2013). Xie 和 Singh(2013) 指出 CD 能够统一许多不同的概念, 包括信仰分布、bootstrap 分布、显著性 (significance) 或 $p$ 值函数、正规化似然函数以及某些情形下的 Bayes 先验和后验分布. 可以说, CD 给出了推断单参数的简单统一范式. 但是, 在试图将其推广到多维参数推断时, 遇到了难以克服的困难, 成为一个公开的问题. 杨振海 (2014) 指出, 用随机估计推断参数就不会存在这一问题, 并在此基础上结合垂直密度表示 (vertical density representation, VDR) 理论, 给出了一个通用的 VDR 检验方法. 虽然随机估计不是为了解决 CD 概念多维参数下的推广问题, 但却为我们提供了一类非常有用的分析工具.

上述统计推断思想和随机估计推断工具潜在的一个重要应用领域是现今流行的各类复杂数据模型的统计推断, 如非平衡、异方差、序列相关和空间面板数据回归模型 (Baltagi, 2013), 纵向数据混合效应模型 (Chen, Dunson, 2003; Vaida, Blanchard, 2005; Bondell, Krishna, Ghosh, 2010; Ni, Zhang, Zhang, 2010; Ibrahim, Zhu, Garcia, Guo, 2011; Fan, Li, 2012) 以及小域估计和预测 (Hall, Maiti, 2006; Jiang, Nguyen, Rao, 2011). 这些数据和模型一般具有复杂、多水平和分层异质性的结构, 不同水平之间以及同一水平内的观测都可能是相关的, 例如空间 (spatial) 数据等. 关于这些数据模型的回归建模、变量选择问题已经受到了广泛关注. 现在研究的热点之一是线性混合效应模型中固定效应和随机效应的选择问题, 并有许多学者进一步研究了广义线性混合效应模型、非线性混合效应模型和非参数、半参数混合效应模型等中的模型选择问题. 如 Bondell, Krishna 和 Ghosh(2010) 考虑了固定和随机效应的联合选择; Ibrahim 等 (2011) 使用了一种新颖的再参数化方法, 将混合效应的选择看成模型中具有很多缺失数据的分组变量选择, 其中的缺失数据代表随机效应; Ni(2010) 等提出了面板数据半参数混合模型中同时进行变量选择和模型估计的双惩罚似然方法. Fan 和 Li(2012) 则主要考虑在线性混合效应模型中选择显著的固定效应. Müller, Scealy 和 Welsh(2013) 综述了线性混合效应模型近期的大量研究, 概括了其中的四种主要方法: 信息准则方法, 如 AIC 或 BIC; 基于惩罚损失函数的压缩方法, 如 Lasso、Fence 方法和 Bayes 技术.

以上变量选择没有关注一个重要问题, 就是由于模型选择步骤中产生的变异

性, 可能给出一个不稳定 (erratic)、具有跳跃性 (jumpy) 的选择后估计. Efron(2014) 则及时地关注了这一问题, 并使用 bootstrap 光滑方法 (也称 bagging) 在考虑变量选择下提供了一种计算标准差和置信区间的一般方法. 我们注意到, CD 和随机估计的研究多是针对较简单经典数据模型的统计推断, 没有对复杂数据统计模型考虑发展合适的统计推断理论和方法, 进而提供相应的置信分布或随机估计理论以及有效的 Meta 分析方案. 因而, 对于实际应用中常见的重复观测数据、复杂面板 (panel) 数据或纵向 (longitudinal) 数据、具有复杂分层结构的时空数据和聚类数据等海量复杂数据, 如何构造出适合的置信分布和随机估计, 开发出相应的 Meta 分析组合技术, 还有许多问题值得研究.

## 1.3　Bootstrap 统计推断

### 1.3.1　Bootstrap 方法简介

Bootstrap 方法是一种为确定估计值精确性的基于计算机的方法, 其基本思想十分简单, 但非常有用.

设 $x=(x_1,\cdots,x_n)$ 是来自分布 $F$ 的一组独立同分布的数据, 分布 $F$ 的均值为 $\mu$, 方差为 $\sigma^2$. 由 $\bar{x}\sim\left(\mu,\dfrac{\sigma^2}{n}\right)$, 可知均值的标准差为 $se(\bar{x})=\sqrt{\dfrac{\sigma^2}{n}}$. 该标准误差常见估计为

$$\widehat{se}(\bar{x})=\sqrt{\frac{s^2}{n}},\quad \text{这里}\quad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{1.3.1}$$

估计的标准误差是一个估计最常用的精确性的度量. 例如, 我们可以粗略地认为 $|\bar{x}-\mu|\leqslant\widehat{se}(\bar{x})$ 有 68%的机会发生, $|\bar{x}-\mu|\leqslant 2\widehat{se}(\bar{x})$ 有 95%的机会发生.

**问题**　设总体 $F\to x=(x_1,\cdots,x_n)$ 是一组独立同分布的数据, 对于总体某个参数 $\theta$ 更复杂的估计量 $\hat{\theta}=T(x)$(如中位数的估计), 如何求 $\hat{\theta}$ 的标准差 $se_F(\hat{\theta})$ 的估计?

**定义 1.3.1** (Bootstrap 方法)　令

$$\hat{F}\to x^*=(x_1^*,\cdots,x_n^*) \tag{1.3.2}$$

是一个容量为 $n$ 的 bootstrap 样本, 而 $\hat{F}$ 是原始样本的经验分布. 根据一个 bootstrap 样本, 可以得到 $\hat{\theta}=T(x)$ 的一个 bootstrap 复制: $\hat{\theta}^*=T(x^*)$. $\hat{\theta}$ 的**标准差 $se_F(\hat{\theta})$ 的 bootstrap 估计**, 就是它的插入 (plug-in) 估计: $se_{\hat{F}}(\hat{\theta})=se_{\hat{F}}(\hat{\theta}^*)$, 也称为 $se_F(\hat{\theta})$ 理想的 bootstrap 估计 (ideal bootstrap estimate).